Medición del Razonamiento Estadístico Descriptivo en Estudiantes Universitarios utilizando la Teoría de Respuesta al Ítem

Karen Yurley Meneses Bautista

Trabajo de Grado para Optar el Título de Matemática

Director
Tulia Esther Rivera Flórez
Magister en Estadística

Universidad Industrial de Santander
Facultad de Ciencias
Escuela de Matemáticas
Bucaramanga
2020

Dedicatoria

A DIOS Uno y Trino, por guiarme y darme la fortaleza para culminar esta meta de ser profesional. Gracias por demostrarme su infinito amor y sabiduría, en las diferentes etapas y pruebas difíciles de mi vida. Cada logro alcanzado, enteramente a Él.

A la Santísima Virgen María por ser protectora e intercesora de mi camino y brindarme consuelo, confianza y esperanza en cada momento que lo necesitaba.

A mi madre Claudia Liliana por su apoyo incondicional y ánimo para alcanzar mis metas. Cada esfuerzo es motivado en sus palabras, gracias por creer en mí.

A mi padre Henry por su esfuerzo y acompañamiento en mi vida. Sin su ayuda no hubiera sido posible culminar esta meta.

A mi hermana Karol por ser incondicional y alentarme en todo momento. Siendo testigo de la realización de mis sueños.

A mi abuelo José por sus enseñanzas y cariño, en este reto universitario fue uno de mis incentivos para culminarlo. Siempre estará en mis recuerdos y corazón.

A mis abuelos, Camilo y Rosa Delia, por su ayuda durante este proceso. Mil gracias.

Agradecimientos

A la profesora Tulia Esther por su paciencia, orientación, confianza y apoyo brindado durante la realización de este proyecto. Gracias por su preocupación, consejos y motivación, muy particular, de animarme a seguir adelante con mis estudios.

A los profesores de la Escuela de Matemáticas que fueron parte fundamental en mi formación académica.

A mis familiares, amigos y compañeros de la carrera, gracias por aportar ese granito de arena durante mi etapa universitaria para que mis logros se cumplieran.

Tabla de Contenido

Pág.

Lista de Tablas

Lista de Figuras

Lista de Apéndices

Resumen

Título: Medición del razonamiento estadístico descriptivo en estudiantes universitarios utilizando la teoría de respuesta al ítem [*]

Autor: Karen Yurley Meneses Bautista [**]

Palabras Clave: Banco de ítems, dominios, nivel de dificultad, nivel de habilidad.

Descripción:

El propósito de este trabajo es aplicar la Teoría de Respuesta al Ítem (TRI) al diseño de un test, el cual pretende medir la capacidad para analizar datos; lo anterior, supone el uso de herramientas estadísticas de tipo descriptivo. La metodología implementada incluyó varias etapas: edición de un banco de ítems, construcción de la primera versión del test, pilotaje, calibración y ajuste, aplicación de la versión final y evaluación de las propiedades psicométricas; y, análisis de resultados.

En cuanto al banco de ítems se logró recopilar una cantidad importante; del total, se pasó a una versión final del test con 27 ítems, acorde al funcionamiento en la prueba piloto, la longitud de la prueba, la participación de todos los dominios de evaluación y la inclusión de diferentes niveles de dificultad. Sobre la calibración del test se logró validar de manera aceptable su confiabilidad, así como los supuestos de unidimensionalidad e independencia local, pese al bajo tamaño de muestra del que se disponía.

Al final se concluyó que a la luz de los datos disponibles, el modelo TRI que obtuvo el mejor ajuste fue el logístico de un parámetro. En cuanto al rendimiento de la muestra de estudiantes, los niveles de habilidad observados se ubicaron en un rango desde -1,6 a 2,09, presentándose gran acumulación en el intervalo de -0,9 a 1 lo que evidencia que en general las habilidades desarrolladas para analizar datos desde un enfoque descriptivo pueden ubicarse en nivel bajo y básico.

[*] Proyecto de Grado.
[**] Facultad de Ciencias. Escuela de Matemáticas. Director: M. Sc., Tulia Esther Rivera Flórez.

Abstract

Title: Measuring descriptive statistical reasoning in university students using item response theory [*]

Author: Karen Yurley Meneses Bautista[**]

Key Words: Bank of items, domains, level of difficulty, skill level.

Description:

The purpose of this paper was to apply Item Response Theory (IRT) to the design of a test that aims to measure the ability to analyze data, which involves the use of descriptive statistical tools. The methodology implemented included several stages: edition of a bank of items, construction of the first version of the test, piloting of the first version of the test, calibration and adjustment, application of the final version of the test and evaluation of the psychometric properties of the test, and analysis of results.

As regards the item bank, a significant number of items were collected, and from this number, a final version of the test was developed with 27 items according to the functioning in the pilot test, the length of the test, the participation of all the evaluation domains and the inclusion of different levels of difficulty. Regarding the calibration of the test, it was possible to validate in an acceptable way the reliability of the test as well as the assumptions of one-dimensionality and local independence despite the low sample size available.

Finally, it was concluded that considering the available data, the IRT model that obtained the best fit was the logistic one parameter. Regarding the performance of the sample of students, the observed skill levels were in a range from -1.6 to 2.09, with a large accumulation in the interval of -0.9 to 1, which shows that in general the skills developed to analyze data from a descriptive approach can be placed in low and basic level.

[*] Undergraduate Project Thesis.

[**] Faculty of Science. Mathematics School. Director: M. Sc., Tulia Esther Rivera Flórez.

Introducción

La investigación en psicología se basa en la recolección de información relacionada con rasgos de personalidad, aptitudes, habilidades o competencias que tienen la particularidad de ser atributos no medibles directamente los cuales suelen ser denominados rasgos latentes. Para tal propósito, desde la psicometría se ha desarrollado todo un marco teórico y metodológico para la construcción de los test y se han planteado diferentes modelos que permiten evaluar la calidad de una medición. En particular, se hace mención a la Teoría de Respuesta al Ítem (TRI) como el recurso base para estimar el error que se comete cuando se mide un dominio o rasgo psicológico.

En la TRI, el propósito es modelar la probabilidad de que un individuo con cierta habilidad responda acertadamente una pregunta de cierto dominio. Estos modelos aportan puntajes en una escala que es comparable aunque los test presentados por los individuos no sean los mismos, tal atributo es posible porque los ítems o preguntas utilizados son previamente calibrados para satisfacer ciertas características estadísticas. Así, para un ítem particular se debe previamente no sólo validar su contenido sino estimar características psicométricas que principalmente tienen que ver con su nivel de dificultad y su poder discriminatorio. En cuanto a la fiabilidad, la TRI incorpora el concepto de precisión local el cual hace referencia a la información que permite evaluar en qué rango de la variable el test discrimina mejor.

En el ámbito educativo mundial, las pruebas estandarizadas se han logrado afianzar como un importante instrumento para medir calidad (logro académico) y soporte para la definición de política educativa no obstante a la controversia que esto genera. En nuestro escenario nacional la experiencia más consolidada la tiene el Instituto Colombiano para la Evaluación de la Educación (ICFES) con las denominadas Pruebas Saber en las que se ha pudo avanzar hasta tener una

estrategia de evaluación longitudinal con puntos de medición desde la básica hasta la educación superior para las áreas de Razonamiento Cuantitativo y lectura crítica que ha permitido implementar estudios de valor agregado y aporte relativo de las instituciones de educación superior; también es de resaltar que el ICFES administra pruebas estandarizadas internacionales entre las que se destaca la del Programa para la Evaluación Internacional de Alumnos (PISA).

Las condiciones actuales suponen grandes retos para las instituciones educativas de nivel superior de índole tanto académica o administrativa, así a nivel mundial se habla de los retos que deberá afrontar la educación, en especial la pública, que van desde el financiamiento, pasando por la cobertura, equidad, investigación, procesos de enseñanza, uso de nuevas tecnologías y acreditación de calidad principalmente. En este sentido, la evaluación viene ganando terreno y es un tema que está en el escenario de discusión, en este aspecto vemos que el modelo evaluativo tradicional de la Universidad Industrial de Santander debe ser revisado y abrir espacios para la incorporación de procesos estandarizados toda vez que si se quiere ampliar cobertura, cada vez las poblaciones atendidas serán más numerosas, como las observadas en el ciclo básico o en ciertos programas de ingeniería con gran afluencia de estudiantes; en particular, este tipo de pruebas permitirían optimizar procesos evaluativos como son: admisión a un programa, evaluación de cursos del ciclo básico, validación por suficiencia y habilitación.

Adicionalmente, el uso de tecnología para acompañar los procesos de evaluación aporta no solo en la optimización de recursos como el tiempo o de logística, también es importante considerar el aporte que hace el soporte computacional en la estimación del nivel de habilidad de un estudiante en cuanto a la calidad en dicha estimación y la posibilidad de ofrecer un informe más completo que vaya más allá de una mera cuantificación numérica poco informativa; así, cada estudiante

puede recibir inmediatamente después del test su retroalimentación de manera individual pero además el consolidado por unidad académica puede ser utilizado como soporte en la toma de decisiones respecto a los procesos de enseñanza.

El presente en este tema, son las pruebas adaptativas computarizadas que están resultando muy útiles cuando se debe evaluar en forma periódica a un número considerable de individuos como sucede en las evaluaciones de logro académico a nivel nacional o en procesos de admisión a una Universidad, también cuando se requiere de una estimación muy precisa del nivel de habilidad de un individuo como sucede en procesos de selección de personal o certificación de competencias laborales.

De otro lado, la relevancia de profundizar la investigación en relación con el razonamiento estadístico es indiscutible dado que nos encontramos ante un nuevo paradigma para describir el mundo y sus fenómenos: el análisis de datos. En particular, Garfield y Chance (2000) definen razonamiento estadístico como la forma en que una persona razona con ideas estadísticas y da sentido a la información estadística; sobre estudios de medición de este tipo de razonamiento, a nivel internacional ha sido un tema de interés y cuenta con importantes referencias entre las que se destacan los tests: *Statistical Reasoning Assessment* (SRA), *Comprehensive Assessment of Outcomes in a First Statistics Course* (CAOS), *Goals and Outcomes Associated with Learning Statistics* (GOALS) y *Basic Literacy in Statistics* (BLIS) según lo reportan Sabbag, Garfield y Zieffler (2018). No obstante a nivel nacional no se encontraron evidencias de estudios en este tema por tanto se justifica profundizar la investigación en esta línea para tener resultados acorde a nuestro contexto.

Las pruebas estandarizadas son instrumentos de medición que permiten evaluar conocimiento o habilidades las cuales han sido uniformizadas en cuanto a los métodos para administrarlas, calificarlas e interpretar los resultados. Entre las ventajas que supone el uso de estas pruebas están la posibilidad de comparación entre individuos o grupos que toman la prueba y la validez y confiabilidad del instrumento.

El origen de las pruebas estandarizadas se ubica en 1905 cuando Binet y Simon formalizaron este concepto realizando un test de coeficiente intelectual basándose en la edad o nivel de educación de los individuos. A pesar de que se desarrollaron investigaciones esporádicas en los años veinte, treinta y cuarenta, formalmente el origen de estas pruebas remonta entre 1950-1960 según Navas (1994), cuando surge la Teoría de Respuesta al Ítem (TRI). La relación entre TRI y pruebas estandarizadas se documenta desde la obra de Lord y Novick (1968), "*Statistical theories of mental test scores*"; en este libro, Birnbaum participo en la elaboración de algunas secciones formulando los fundamentos de la TRI y la forma de ordenar estadísticamente los diferentes participantes en una misma escala de medida sin importar que hayan presentado test diferentes, un tratamiento matemático asequible que allana el camino hacia la estimación de parámetros.

Posteriormente, Lord (1970) planteó las bases para realizar pruebas estandarizadas por medio de la Teoría de Respuesta al ítem (TRI), investigación que influyó en el desarrollo de la prueba adaptativa *Armed Services Vocational Aptitude Battery* (ASVAB) que se dio con el convenio del ejército estadounidense y la Universidad de Minnesota para elaborar una prueba de selección de aspirantes a integrar el ejército nacional. Antes de ponerla en práctica transcurrieron diez años realizando ensayos y analizando las ventajas de esta prueba adaptativa en dicha selección de personal, finalmente a mediados de los años 80 se aplicó oficialmente por primera vez la prueba del ASVAB.

También en la década de 1970, ya se contaba con las primeras computadoras con capacidad suficiente y accesibilidad para ser utilizadas en pruebas estandarizadas pero sin la agilidad necesaria para tratar con las operaciones matemáticas requeridas por la teoría de respuesta al ítem; a pesar de esta dificultad se crearon lo que conoce hoy en día como pruebas adaptativas informatizadas (TAI) para medir aptitudes escolares que fueron propuestas por diferentes autores y que se referencian en Olea y Ponsoda (2002).

Después de unas décadas, las pruebas adaptativas computarizadas han alcanzado gran crecimiento en su aplicación y algunas pruebas que se realizaban de forma tradicional fueron transformadas al formato CAT. La anterior descripción muestra que, aunque el planteamiento de la Teoría de Respuesta al Ítem fue dada en los años 60 su implementación definitiva se dio por Lord (1980) con la publicación de su libro *Applications of ítem response theory to practical testing problems.* La demora se pudo originar en que su desarrollo fue motivado más para resolver problemas técnicos relacionados con la construcción de tests y de la estadística matemática involucrada. Hoy día esta teoría sigue vigente y ganando gran popularidad, principalmente en estudios de medición en psicología y educación.

En el escenario actual, a nivel internacional se encuentran numerosas aplicaciones de este tipo de pruebas, Olea y Ponsoda (2002) en su tesis mencionan algunas pruebas principalmente orientadas a evaluar razonamiento verbal, razonamiento cuantitativo como es el Graduate Record Exam (GRE), test sobre el dominio de un idioma, el Test of English as a Foreign Language (TOEFL), pruebas de ingreso a centros educativos ya sea en educación básica o universitaria por ejemplo, Graduate Management Admissions Tests (GMAT), Law School Admission Test COMPASS placement tests, el NWEA o el sistema CARAT, también existen múltiples test para selección de personal (como el CAT-ASVAB o el MICROPAT con el fin de evaluar los

conocimientos de pilotos y controladores aéreos), o test que evalúa el conocimiento matemático en adultos como la prueba MATHCAT.

En cuanto al desarrollo de tests en el campo de la Educación Estadística, DelMas, Garfield, Ooms y Chance (2007) desde 1999 se enfocaron en direccionar los retos de la evaluación en esta área hacia el diseño de instrumentos confiables, válidos, prácticos y accesibles. Como resultado de esta iniciativa hoy día se cuenta con un sitio web llamado ARTIST donde se dispone de un banco de ítems de acceso libre. En este sentido, el test CAOS (Comprehensive Assessment of Outcomes in Statistics disponible en https://app.gen.umn.edu/artist/) permite evaluar el razonamiento estadístico después de un primer curso de estadística centrándose en la cultura estadística y la interpretación de conceptos. El instrumento fue aplicado a una muestra de 23.645 estudiantes y sometido a pruebas de análisis de contenido y confiabilidad las cuales permitieron concluir que dicho test es un instrumento válido para medir importantes resultados de aprendizaje en un primer curso de Estadística. Posteriormente, Ziegler (2014) desarrolló otro instrumento de evaluación denominado Basic Literacy in Statistics (BLIS) cuyo propósito fue medir la cultura estadística, entendida como la habilidad de leer, entender y comunicar información estadística ganada tras superar un primer curso de Estadística que enfatizó el uso de métodos de simulación.

En relación a la evaluación estandarizada en razonamiento estadístico a nivel de Colombia, se encontraron referencias de las pruebas administradas por el ICFES y el examen de Ciencias Básicas (EXIM) administrado por la Asociación Colombiana de Facultades de Ingeniería (ACOFI). Cabe aclarar que dichos test incluyen dentro de su prueba de razonamiento cuantitativo en el primer caso y en la de matemáticas en el segundo caso, algunas preguntas en probabilidad y análisis estadístico de datos. A continuación, se resumen los trabajos que se encontraron donde se observa una implementación practica de pruebas estandarizadas ajustadas al modelo Rasch.

- Morgado y Neusa (2011) realizaron un análisis de los resultados en las Olimpiadas Regionales de Matemáticas UIS, basado en: sedes regionales de la Universidad, niveles (básico, medio y avanzado) y áreas. Estas Olimpiadas poseen cinco fases de ejecución de las cuales los autores seleccionaron tres: clasificatoria, selectiva y final, para analizar descriptivamente los ítems. El estudio fue realizado con los datos obtenidos en los años 2009 y 2010, y para su ejecución utilizaron una muestra de estudiantes de secundaria de colegios públicos y privados que presentaron las olimpiadas en las diferentes sedes de la UIS. Otros objetivos de su investigación fueron analizar los niveles de habilidad de los estudiantes, la dificultad de los ítems, los errores cometidos por parte de los individuos en esta prueba y el ajuste de los datos al modelo Rasch.

  Se realizó la prueba estadística t la cual permitió determinar que el comportamiento de los resultados en los dos años fue similar en el nivel básico y medio en las diferentes sedes durante la fase clasificatoria; y se ejecutó la prueba chi cuadrado para la bondad de ajuste en la fase clasificatoria que evidenció que los individuos del nivel básico y avanzado durante los dos años pudieron responder al azar o adivinando los ítems, esto comprueba los problemas que se presentan en el desarrollo de las olimpiadas.

  Tras analizar los resultados de la prueba los autores determinaron que se presentó un buen ajuste de los datos en el modelo empleado en la fase clasificatoria y se realizó una estimación adecuada a los parámetros de dificultad de los ítems, pero no se pudo establecer una estimación de los niveles de habilidad de los estudiantes debido a que la muestra era grande y la cantidad de ítems era pequeña en esta fase de las olimpiadas, esta situación también se presentó para los ítems de la fase clasificatoria unida con la fase selectiva. Finalmente, al unir los resultados de las tres fases en su análisis se obtuvo un buen ajuste

al modelo, lo que permitió estimar los niveles de habilidad de los individuos y conllevaron a determinar que los ítems implementados en las fases estudiadas son adecuados para medir las habilidades de los estudiantes que lleguen a la etapa final de las olimpiadas.

- Barajas y Esparza (2010) construyeron e implementaron un test compuesto por veintitrés ítems con el objetivo de estimar la habilidad algebraica de individuos que ingresaban a primer semestre en las carreras de ciencias, ingenierías y matemáticas de la Universidad Industrial de Santander en el segundo semestre de 2009, aplicaron el test a una muestra de 319 estudiantes; se planteó identificar los errores algebraicos más comunes en los estudiantes, adicionalmente se estudió el ajuste de los datos al modelo Rasch, se realizó la estimación del parámetro de dificultad de los ítems del test y del nivel de habilidad de los individuos.

  Al implementar el test y al analizar los resultados obtuvieron varias conclusiones. En primera estancia pudieron constatar con respecto a su objetivo que los niveles de habilidad de los estudiantes estuvieron entre -2,30 y 2,75 con una distribución aproximadamente normal, al realizar una comparación de los niveles de habilidad de los individuos por carreras determinaron que los estudiantes que ingresaron a las ingenierías tenían un mejor desempeño. Adicionalmente, que los ítems donde más se detectaron errores por parte de los estudiantes eran aquellos en que debían simplificar fracciones y aplicar la propiedad distributiva; el parámetro de dificultad de los ítems obtuvo valores entre -2 a 2,78 estableciendo que los ítems de mayor dificultad eran aquellos donde se hacía operaciones con fracciones y los ítems de menor dificultad eran los que tenían operaciones con funciones lineales, este análisis de la distribución del parámetro de dificultad les permitió proponer nuevos ítems para mejorar el test de tal forma que fortalezca el banco de ítems

para construir subtests que midan niveles de habilidad algebraicos. Finalmente, los datos presentaron un buen ajuste al modelo Rasch.

En este trabajo nos proponemos aplicar los elementos teóricos de la TRI para el diseño de un test que pueda ser utilizado posteriormente en procesos de evaluación a gran escala, el dominio a evaluar sería el análisis estadístico de datos dada su relevancia actual para todo tipo de profesional y en consecuencia la facilidad para conseguir una muestra de individuos para aplicar las pruebas requeridas. Bajo la perspectiva de implementar en el largo plazo una prueba adaptativa computarizada, este trabajo de grado cubrirá solo las primeras etapas de una metodología en tal sentido las cuales son: el diseño de un primer banco de ítems y su calibración.

Finalmente es válido mencionar que este tipo de trabajos contribuyen a fomentar la investigación en temas de educación bajo un enfoque cuantitativo, línea de trabajo de gran relevancia en la actualidad y que puede ser de gran interés para diferentes unidades académicas de nuestra Universidad y del país.

1. Objetivos

1.1 Objetivo General

Medir el razonamiento estadístico descriptivo en estudiantes universitarios mediante la Teoría de Respuesta al Ítem.

1.2 Objetivos Específicos

Diseñar un test que permita estimar el nivel de competencia en temas relacionados con análisis estadístico de datos.

Calibrar el test acorde a un modelo de Teoría de Respuesta al Ítem.

2. Marco Teórico

A continuación se presentan los elementos teóricos que desde la psicometría se plantean en relación con la evaluación bajo el enfoque de la TRI. Para empezar presentamos la terminología propia de esta área que se complementa con el vocabulario propio del argot estadístico:

Rasgo: hace referencia al dominio o dimensión a evaluar, puede ser de tipo individual o compuesto cuando reúne varias habilidades y conocimientos. En las aplicaciones de la TRI se quiere medir habilidad (pruebas académicas) o rasgos de personalidad (test psicológicos).

Constructo latente o Rasgo latente: variable no observada que da cuenta de la aptitud de un examinado, puede denominarse también como atributo, rasgo de interés, nivel de competencia o nivel de habilidad que evalúa la prueba.

Ítem o reactivo: pregunta que conforman el test o prueba.

Fiabilidad o confiabilidad del test: es una propiedad psicométrica que se refiere a la consistencia de las puntuaciones obtenidas por los individuos cuando son examinados con el mismo test en diferentes ocasiones.

Error de medición: distorsión en la estimación del rasgo latente que subyace al modelo, surge porque al responder un test hay influencia de variables que pueden afectar las respuestas y que no pueden ser controladas.

2.1. Metodología para creación de un Test

Los elementos teóricos que fundamentan la predicción que puede hacerse bajo un enfoque de evaluación basado en la teoría de respuesta al ítem fueron tomados de las siguientes fuentes Nava

(1994), Muñiz (1998), Olea y Ponsoda (2002), Argibay (2006), Attorresi, Lozzia, Abal y otros (2009), Pérez (2011), Leenen (2014), Goforth (2015), Hidalgo y French (2016), Carvajal, Méndez y Torres (2016), EMAR (2018), Reckase (2018), Paek y Cole (2020).

2.1.1. Banco de Ítems. Es un conjunto de preguntas (enunciado y opciones de respuesta), información psicométrica (parámetros estimados a los ítems) e información complementaria como puede ser el contenido que mide, tasa de exposición en aplicaciones anteriores, distribución de respuestas en los distractores propuestos, etc.

Para su diseño, la primera consideración que hay que hacer es analizar el propósito de la prueba. Principalmente se considera que puede ser: medir rendimiento máximo, medir rendimiento típico (usual para rasgos de personalidad) o clasificar una población en subgrupos; una vez se tiene claro el propósito de la prueba el siguiente paso es determinar características técnicas como:

- *Tipo de información*: decidir si los enunciados serán sólo de tipo verbal o incluirán información gráfica, actualmente debe considerarse la posibilidad de una presentación dinámica de los enunciados.
- *Tipo de pregunta*: por ejemplo, cuando se pretende medir el rendimiento máximo (conocimientos o rasgos intelectuales) lo usual ha sido utilizar preguntas de selección múltiple, aunque se advierte la necesidad de preguntas abiertas pensando en la evaluación de áreas como matemáticas o programación.
- *Tamaño del banco*: se determina dependiendo del algoritmo de selección de ítems que se establezca y de aspectos adicionales como si hay categorías de contenido en la prueba, amplitud del rango de los niveles de rasgo (habilidad) a medir, y el número de aplicaciones que se vayan a hacer en el futuro para proveer la rotación de las preguntas.

- *Calibración del banco de ítems:* este proceso se implementa en relación a un modelo concreto de la TRI. A este respecto, Olea y Ponsoda (2002) sugieren incluir aquí aspectos como decidir el tamaño muestral que se requiere para la calibración de las preguntas, elección del modelo TRI más apropiado, determinar si se va a establecer un diseño de anclaje y equiparación (para pruebas con alto número de ítems se opta por dividir el test en secciones que se aplican a muestras diferentes), estimar los parámetros del modelo y comprobar la bondad de ajuste del modelo TRI seleccionado (Unidimensionalidad e invarianza), medir indicadores psicométricos clásicos (índice de discriminación, validez de contenido, confiabilidad, etc.) y decidir sobre ítems a eliminar.
- *Mantenimiento y renovación de banco de ítems*: hace referencia a acciones como: eliminar ítems con propiedades psicométricas inadecuadas, estudio de la tasa de exposición de una pregunta, renovación y ampliación del banco de ítems, pilotaje de nuevos ítems, recalibrar los ítems originales y revisar a partir de datos nuevos la estimación de los parámetros del modelo en relación con la estimación inicial.
- *Otras consideraciones*: A parte del tipo de pregunta, son temas de investigación actual el numero óptimo de opciones que debe tener una pregunta de selección múltiple y el impacto que tiene el orden de aparición de las opciones de respuesta, metodologías para la construcción de los ítems, optimización del tamaño de la prueba, producción de indicadores adicionales al puntaje numérico que complementen la evaluación (tiempo de respuesta, éxito o fracaso según el tipo de pregunta, etc.) entre otros.

Los ítems que componen un test pueden ser dicotómicos o politómicos; los ítems dicotómicos también se conocen como binarios y se denominan así cuando existe solamente dos opciones de respuesta, por ejemplo, correcto o incorrecto, sí o no, etc. Los ítems politómicos son los que poseen

más de dos opciones de respuesta, por ejemplo, escala de respuesta estilo Likert, calificación crediticia parcial o respuestas de opción múltiple.

2.2. Elementos teóricos básicos en medición de test

A diferencia de los procesos de medición de atributos físicos como puede ser un área o una estatura, el test no es un instrumento que produzca una medición directa, es un experimento en el que produce varias respuestas a partir de las cuales se inferirá la medición deseada, en este caso la aptitud del examinado. Para este propósito se cuenta con un amplio desarrollo teórico que incluye dos enfoques básicos: la Teoría Clásica de los Test (TCT) y la Teoría de Respuesta al Ítem (TRI) que han sido planteados desde la psicometría.

Inicialmente, la TCT que también se denomina *modelo de la puntuación verdadera* o *teoría del error de medición*, se planteó a principios del siglo XX introduciendo expresiones como *puntuación verdadera* y *puntuación observada* con sus respectivas diferencias. En la segunda mitad del siglo XX se dan los cimientos de la TRI que consistía en analizar las respuestas en una prueba por medio de los ítems en vez de hacerlo con el resultado total de la medición.

Actualmente han avanzado las investigaciones y el desarrollo teórico de la TRI ha tomado gran importancia durante las últimas décadas desde su aparición. Su crecimiento ha ido de la mano con la tecnología, por ello hoy en día cuenta con un grupo grande de modelos psicométricos, los cuales comparten el objetivo de obtener información probabilística de la respuesta de la persona en cada ítem, para llevarlo a cabo buscan relacionar por medio de un modelo matemático las propiedades no observables de los ítems en una prueba y de las personas que los contestan.

La TCT se sigue utilizando primordialmente en el estudio de los resultados de un test a pesar de que la TRI es más avanzada y extensa en teoría, esto es debido a que aún existen áreas donde son insuficientes los avances de las investigaciones sobre el análisis de información que proporciona los instrumentos de evaluación con base a la TRI.

2.2.1. Modelo Rasch. Se utiliza para medir un fenómeno latente, no observable directamente, partiendo de las estimaciones de los parámetros de las personas y de los ítems, se puede expresar la probabilidad de respuesta a un ítem *i* con dificultad b_i por parte de un sujeto *j* con habilidad θ_j mediante la ecuación [1], asumiendo que todos los ítems tienen el mismo parámetro de discriminación.

$$P(X_{ij} = 1 \mid \theta_j) = \frac{e^{(\theta_j - b_i)}}{1 + e^{(\theta_j - b_i)}} \quad [1]$$

Donde:

b_i*: Parámetro de dificultad del ítem i*

θ_j*: Rasgo o aptitud latente del individuo j*

X_{ij}*: Respuesta del sujeto j al ítem i*

2.2.2. Teoría de Respuesta al Ítem. La Teoría de Respuesta al Ítem (TRI) tuvo sus inicios en los años cincuenta debido a los problemas y limitaciones que contenía la TCT. Sin embargo Navas (1994) ubica sus antecedentes en la década de los años veinte con los trabajos realizados por Thurstone (1928) y en las décadas de los treinta y cuarenta con los aportes por parte de Richardson (1936), Lawley (1943, 1944) y Tucker (1946).

Posteriormente, Lord y Novick (1968) en el libro *Statistical Theories of Mental Test Scores* dieron su respaldo a la TRI, este suceso hizo que aumentaran los trabajos, investigaciones y

aplicaciones sobre las TRI, e incluso monografías, libros y publicaciones de artículos en revistas importantes para el área, al tiempo que se desarrollan métodos de estimación de parámetros (Wright y Panchapakesean, 1969; Lord, 1974; Bock, 1972) y surgen modelos para distintos formatos de respuesta como: modelos de respuesta guardada (Samejima, 1969), modelo de respuesta continua (Samejima, 1972) o el modelo de respuesta nominal Bock (1972).

La Teoría de Respuesta al Ítem tiene un panorama diferente con respecto a otras teorías, esta se centra en la información que ofrece un determinado ítem, se podría decir que la TRI es un modelo estadístico que estima la probabilidad de respuesta a un ítem en función de un parámetro específico del ítem (su dificultad) y el nivel de habilidad que presenta el individuo que lo responde. También es importante resaltar que el objetivo de la TRI es la elaboración de herramientas de medición con propiedades invariantes entre poblaciones, es decir, si dos individuos poseen la misma habilidad tienen la misma probabilidad de responder acertadamente el mismo ítem. Igualmente, si dos ítems son igualmente difíciles, la probabilidad de responderlo acertadamente es la misma para dos individuos con el mismo nivel de habilidad.

2.2.2.1. Modelos logísticos.

- **Modelo de Respuesta Logístico de un parámetro (1P).** Este modelo incluye como único parámetro la dificultad del ítem (***bi*)** ya que asume que la discriminación **(a)** es igual para todos los ítems, es decir $a_i = a$, como se evidencia en la ecuación [2], se debe tener en cuenta que cuando este último parámetro es 1 se da la ecuación característica del modelo Rasch.

$$P(X_{ij} = 1 \mid \theta_j) = \frac{e^{a(\theta_j - b_i)}}{1 + e^{a(\theta_j - b_i)}} \quad [2]$$

Donde:

a: Parámetro de discriminación de los ítems

b_i: Parámetro de dificultad del ítem i

θ_j: Rasgo o aptitud latente del individuo j

X_{ij}: Respuesta del sujeto j al ítem i

- **Modelo de Respuesta Logístico de dos parámetros (2P).** Este modelo tiene dos parámetros para describir las características psicométricas de un ítem: la dificultad del ítem (b_i) y la discriminación del ítem (a_i) como se muestra en la ecuación [3], establecido por Birnbaum (1957). Más adelante en la sección 4.2.2.3 se describe con detalle el papel de cada uno de los parámetros de los modelos presentados. En el ajuste del modelo 2P usando el paquete "ltm" de R se asume una distribución normal para la distribución poblacional de θ usando el método de máxima verosimilitud.

$$P(X_{ij} = 1 \mid \theta_j) = \frac{e^{a_i(\theta_j - b_i)}}{1 + e^{a_i(\theta_j - b_i)}} \quad [3]$$

Donde:

a_i: Parámetro de discriminación al ítem i

b_i: Parámetro de dificultad del ítem i

θ_j: Rasgo o aptitud latente del individuo j

X_{ij}: Respuesta del sujeto j al ítem i

- **Modelo de Respuesta Logístico de tres parámetros (3P).** En este caso el modelo cuenta con tres parámetros: dificultad del ítem, discriminación del ítem y el parámetro de pseudo-

azar del ítem. Este modelo fue introducido por Birnbaum (1968) y se plantea en la ecuación [4].

$$P(X_{ij} = 1 \mid \theta_j) = c_i + (1 - c_i)\frac{e^{a_i(\theta_j - b_i)}}{1 + e^{a_i(\theta_j - b_i)}} \quad [4]$$

Donde:

a_i: *Parámetro de discriminación al ítem i*

b_i: *Parámetro de dificultad del ítem i*

c_i: *Parámetro de pseudo-azar al ítem i*

θ_j: *Rasgo o aptitud latente del individuo j*

X_{ij}: *Respuesta del sujeto j al ítem i*

El parámetro nuevo acá es c_i que se llama parámetro del seudo-azar porque representa la probabilidad en los ítems de opción múltiple de que un individuo con un nivel de habilidad bajo conteste correctamente un ítem difícil, lo que hace suponer que lo hizo adivinando o por azar.

El resultado más importante de un modelo TRI es la Curva Característica de un Ítem (CCI) que es la representación gráfica de la función que describe la probabilidad de responder exitosamente un ítem a partir del nivel de habilidad del individuo. Este recurso es el centro del análisis de las propiedades de los ítems. La representación gráfica en la Figura 1 muestra un ejemplo típico de una CCI, como se ve ésta curva describe un comportamiento creciente, es decir, la probabilidad de dar una respuesta correcta a un ítem aumenta con el aumento del nivel de habilidad.

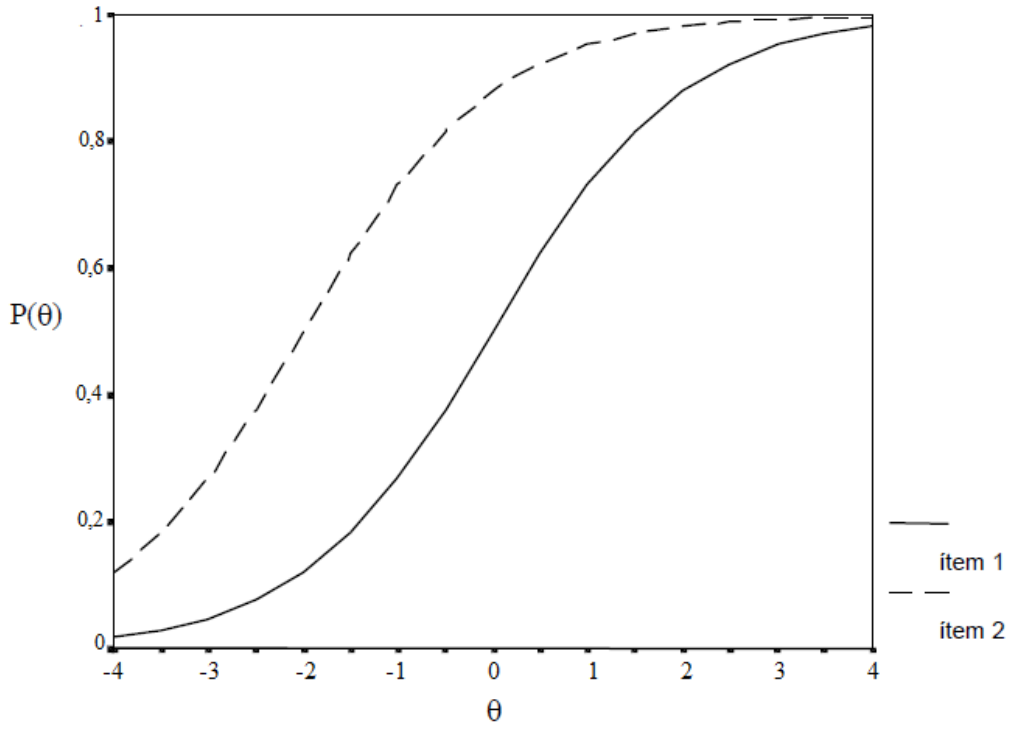


Figura 1. Curva Característica de dos ítems. Adaptada de Olea y Ponsoda (2002). *Test Adaptativos Informatizados.* Madrid.

Cuando son ítems de habilidades, lo fundamental es la respuesta correcta y en los ítems de personalidad es la opción que muestra un nivel mayor de rasgo en el individuo. Cada combinación de parámetros genera una CCI propia, siendo característica su forma de S la cual implica un cambio creciente en la probabilidad de acierto conforme aumente el nivel de habilidad. En la Figura 1 se presentan dos ítems con diferente nivel de dificultad, el ítem 1 con una dificultad de 0 (curva más sesgada a la derecha) y el ítem 2 con un parámetro de dificultad de valor -2 la representación de su curva característica muestra una tendencia de facilidad.

En una etapa posterior del análisis, la información que reflejen estas gráficas contribuye a tomar decisiones como son la eliminación de ítems que no aportan información significativa sobre el rasgo de interés, seleccionar ítems que aportan mayor discriminación en el rasgo e identificar ítems

prejuiciosos, esto es, que son más discriminativos en ciertos grupos de la población, técnicamente esto tiene que ver con la función diferencial del ítem (DIF).

2.2.2.2. Estimación de Parámetros. Cuando se implementa un modelo a los datos su finalidad es obtener estimaciones para los parámetros de ítems y personas por medio de un análisis de las respuestas aportadas por una muestra. Existen varios métodos de estimación para los modelos TRI, entre los que se destacan el de método de máxima verosimilitud y el método bayesiano. El método de máxima verosimilitud busca los valores que hagan máxima la probabilidad de las respuestas observadas, para ello considera diferentes valores estimados para el rasgo latente (θ_j), a partir de la expresión [5]:

$$L(u \mid \theta) = \prod_{j=1}^{n} P_j^{u_j} Q_j^{1-u_j} \quad [5]$$

Donde *u* es el vector de respuestas a los ítems, P_j es la probabilidad de acertar el ítem *j* dado un nivel de rasgo θ y Q_j es la probabilidad de fallarlo. La primera vez que se aplica el test a una muestra se estiman los parámetros de los ítems lo que se denomina calibración del test y al tiempo también se deben estimar el parámetro θ para cada individuo, en las aplicaciones posteriores solo se deberán estimar los niveles de rasgo de los individuos, es decir, los θ_j. Como se ve el problema de estimación es complejo dado el número de parámetros que se desconocen razón por la cual se requiere un gran tamaño de muestra.

La solución analítica de este problema consiste en encontrar la solución al sistema de ecuaciones que se obtiene al igualar a cero la derivada parcial de L descrita en [6] respecto a cada parámetro, desafortunadamente este sistema no tiene solución analítica por lo cual se debe recurrir

a métodos numéricos. En relación a este proceso Olea y Posonda (2002) advierten que en el caso de individuos con patrón de respuesta constante, es decir que todas sus respuestas son aciertos o que todas sus respuestas son fracasos, dicho comportamiento conduce a que no haya un máximo en la función de verosimilitud por lo cual en estos casos se debe optar por un procedimiento de estimación bayesiano que permita obtener estimaciones finitas en estos casos.

$$P(\theta|u) = \frac{g(\theta)L(u|\theta)}{L(u)} \propto g(\theta)L(u|\theta) \text{ [6]}$$

Donde $g(\theta)$ es la distribución a priori de $\theta, L(u|\theta)$ es la función de verosimilitud y L (u) es la verosimilitud del patrón de respuestas. En nuestro caso, el análisis de datos se hará a través del paquete ltm del software R el cual utiliza como método de estimación máxima verosimilitud marginal (MML).

2.2.2.3. Caracterización de los Parámetros a través de una CCI. Para ilustrar cómo interpretar y asignar valores a partir de una CCI utilizaremos los ejemplos citados en Rojas y otros (2004). La primera vez que se aplica un test a una muestra de individuos se estiman los parámetros de los ítems, proceso que se denomina calibración, también se estiman los parámetros θ de los individuos. Así una vez calibrado el test, el problema es estimar los niveles de rasgo de los individuos, problema de índole estadística que consiste en hallar la verosimilitud de un patrón de respuesta observado para cada nivel θ posible.

Para ilustrar cómo hallar el nivel θ que le corresponde a un individuo, Olea y Posonda (2002) utilizan el siguiente ejemplo. Para un test que tiene dos únicos ítems denominados P_1 y P_2, una persona acierta al primero y falla en el segundo, para hallar el nivel θ en este caso hallaríamos la verosimilitud de este patrón de respuesta para cada nivel θ posible (tomando un rango de -4 a +4),

es decir, para cada θ posible se calcularía $L = P_1\,(1 - P_2)$ donde P_i corresponde a la probabilidad de éxito del ítem *i* obtenida a partir del modelo en consideración. Para los patrones de respuesta mencionados en el ejemplo observamos en la Figura 2 que el nivel de habilidad (θ) que cumple con las condiciones dadas es cuando vale -1 y ahí L tiene un valor máximo**.** La gráfica representa los resultados de L para cada valor θ.

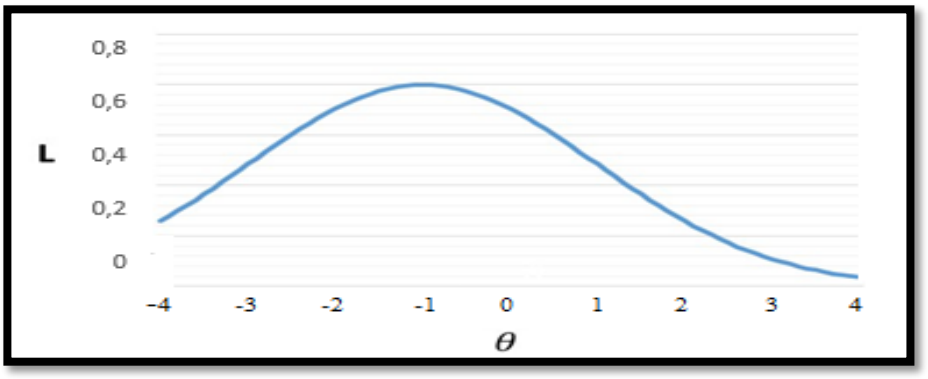


Figura 2. Estimación de L para cada nivel de rasgo. Adaptada de Olea y Posonda (2002). *Test Adaptativos Informatizados.* Madrid.

Parámetro de Discriminación (a_i). Magnitud del cambio en la probabilidad de acertar un ítem conforme varía el nivel de habilidad. Es una medida de la capacidad que posee el ítem de diferenciar entre un individuo hábil y uno que es menos hábil. Gráficamente su valor es proporcional a la pendiente de la recta tangente a la CCI en el punto de inflexión de la curva[1].

[1] La demostración se obtiene usando la ecuación característica [3]. Tenemos que a_i es proporcional a la pendiente de la CCI en el valor $\theta_j = b_i$ y que un punto de inflexión se da cuando la probabilidad de acertar el ítem es 0.5, así

$$0.5 = \frac{e^{a_i(\theta_j - b_i)}}{1 + e^{a_i(\theta_j - b_i)}} \Rightarrow 0.5 + 0.5\,e^{a_i(\theta_j - b_i)} = e^{a_i(\theta_j - b_i)} \Rightarrow 0.5 = e^{a_i(\theta_j - b_i)} - 0.5\,e^{a_i(\theta_j - b_i)} \Rightarrow 0.5 = 0.5\,e^{a_i(\theta_j - b_i)} \Rightarrow 1 = e^{a_i(\theta_j - b_i)} \Rightarrow In(1) = In(e^{a_i(\theta_j - b_i)}) \Rightarrow$$

$$0 = a_i(\theta_j - b_i) \Rightarrow 0 = a_i\theta_j - a_i b_i \Rightarrow a_i\theta_j = a_i b_i \Rightarrow \theta_j = b_i \blacksquare$$

A continuación, se ejemplariza este parámetro con tres distintos índices de discriminación que permiten distinguir unos comportamientos muy diferentes:

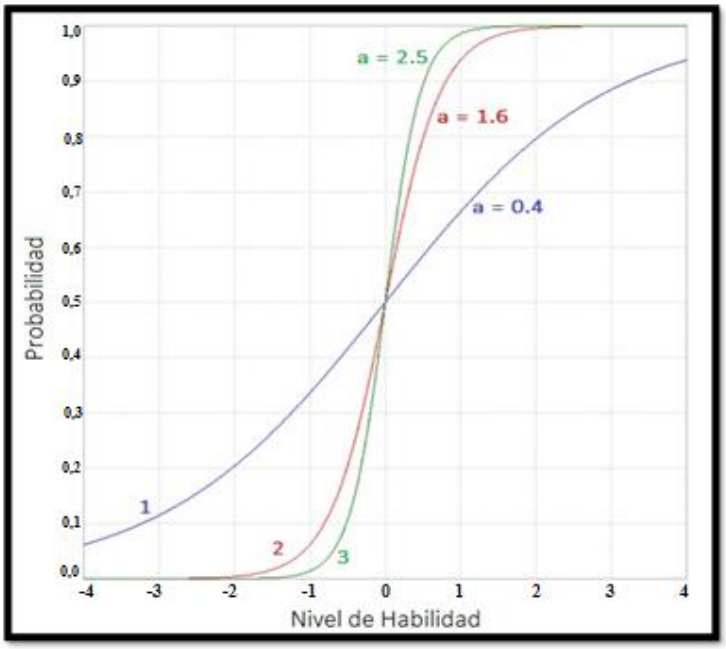


Figura 3. CCI Correspondiente a Preguntas con Diferentes índices de Discriminación. Adaptada de Rojas y otros (2004). *Curso de UML Multiplataforma Adaptativo Basado en la Teoría de Respuesta al ítem. Revista Ingeniería Informática, 10*.

Según la gráfica anterior, la pendiente de la curva número 2, es la más alta, esto indica que este ítem es el más discriminatorio de los tres, ya que, para niveles bajos de habilidad, $\boldsymbol{a}$ menor que -1, es imposible que se responda correctamente a este ítem, contrario a lo que sucede en la parte derecha de la escala, donde para niveles altos de habilidad, es casi seguro una respuesta correcta. En cambio, una pendiente menor como la curva número 1 establece que este ítem es el menos discriminatorio.

En una implementación práctica es importante analizar los ítems y verificar si los ítems funcionan de la forma indicada, de lo contrario ver si se necesita un ajuste o ser reemplazados por otros. Para Hidalgo y French, (2016), un ítem funciona bien si discrimina a los individuos evaluados en todos los niveles de $\boldsymbol{\theta}$, es decir, tener un valor de a (índice de discriminación)

relativamente alto y que se utilicen todas las opciones de respuesta propuestas. Los autores referencian a Baker (2001) quien propone un criterio para saber si los ítems funcionan correctamente, ésta se basa en establecer criterios para el índice de discriminación ($\boldsymbol{a_i}$), por ejemplo, si $\boldsymbol{a_i} < 0.65$ el ítem no cumple con el umbral mínimo y debe ser revisado por un experto en el tema, si $\boldsymbol{a_i} > 1.34$ el ítem tiene un nivel elevado de funcionamiento y si $\boldsymbol{a_i} > 1.69$ se considera que el ítem tiene un nivel de funcionamiento muy elevado. De todas formas, se recomienda hacer una evaluación que considere los resultados estadísticos y el contenido de los ítems.

Parámetro de Dificultad ($\boldsymbol{b_i}$). Es el nivel de dificultad del ítem que se relaciona con la posición de éste en la escala de aptitud, suele identificarse como un parámetro de localización del ítem. En una representación gráfica corresponde a la abscisa (nivel de habilidad) del punto de máxima pendiente [2], a partir de una gráfica basta con ubicar para qué nivel de habilidad la probabilidad es 0.5. La Figura 4 muestra como lucen tres preguntas con índices de dificultad diferentes.

Al analizar la Figura 4 se puede deducir que la curva número 2 corresponde al ítem más fácil, ya que con niveles bajos de habilidad en este caso mayores a -2 se tienen altas probabilidades de

[2] La demostración se obtiene usando la ecuación característica [1 ó 2]. Tenemos que la dificultad es un parámetro de localización del ítem que representa la posición de la CCI en relación al nivel de habilidad necesario para obtener una probabilidad de acierto $P(\theta_j) = \frac{(1+c)}{2}$, donde se obtiene su máxima pendiente y *c* es una constante, así,

$$\frac{(1+c)}{2} = \frac{e^{(\theta_j - b_i)}}{1 + e^{(\theta_j - b_i)}} \Rightarrow \frac{(1+c)}{2} + \frac{(1+c)}{2} e^{(\theta_j - b_i)} = e^{(\theta_j - b_i)} \Rightarrow$$

$$\frac{(1+c)}{2} = e^{(\theta_j - b_i)} - \frac{(1+c)}{2} e^{(\theta_j - b_i)} \Rightarrow \frac{(1+c)}{2} = \frac{(1-c)}{2} e^{(\theta_j - b_i)} \Rightarrow \frac{1+c}{1-c} = e^{(\theta_j - b_i)} \Rightarrow$$

$$In\left(\frac{1+c}{1-c}\right) = In\left(e^{(\theta_j - b_i)}\right) \Rightarrow In\left(\frac{1+c}{1-c}\right) = \theta_j - b_i \Rightarrow In\left(\frac{1+c}{1-c}\right) + b_i = \theta_j \ \blacksquare$$

acertarlo, en cambio la curva número 3 representa el ítem más difícil en comparación con los otros, ya que se necesita un alto nivel de habilidad para ser respondido correctamente.

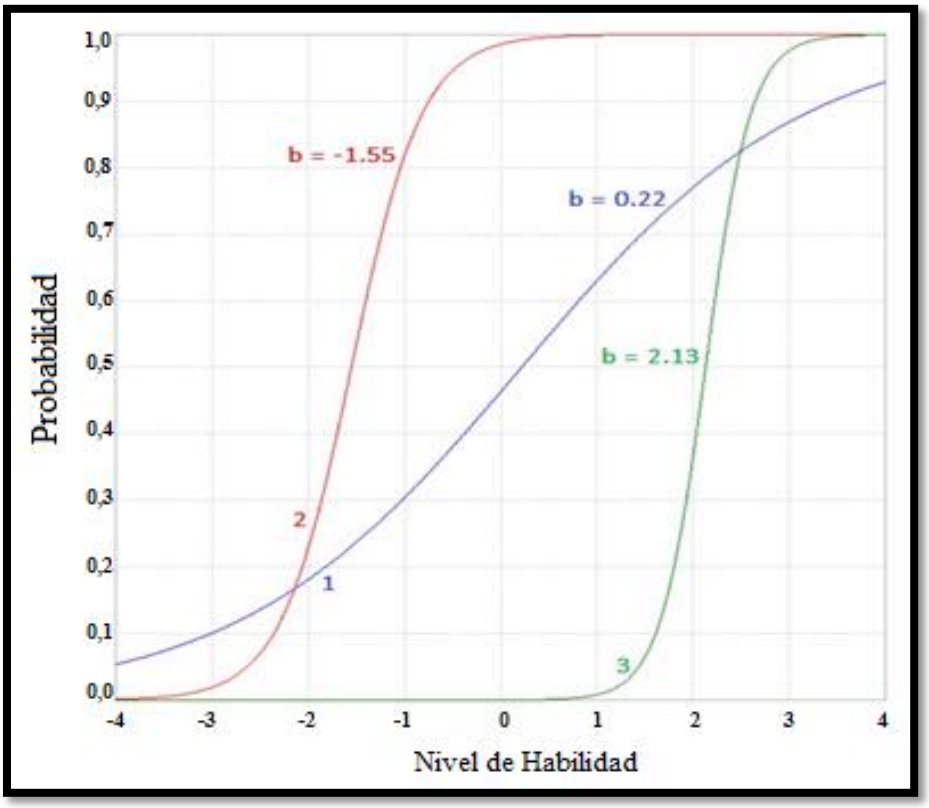


Figura 4. CCI Correspondiente a Preguntas con Diferentes Índices de Dificultad. *Curso de UML Multiplataforma Adaptativo Basado en la Teoría de Respuesta al ítem*. *Revista Ingeniería Informática, 10.*

Parámetro de Pseudo - Azar (c_i). Indica la posibilidad de que un individuo pueda acertar al ítem por azar. En la práctica se asume que es el alumno que posee muy baja habilidad y contesta adivinando dando una respuesta correcta.

Para ilustrarlo gráficamente se consideran tres valores distintos para este parámetro de pseudo azar de tres preguntas diferentes (ver Figura 5).

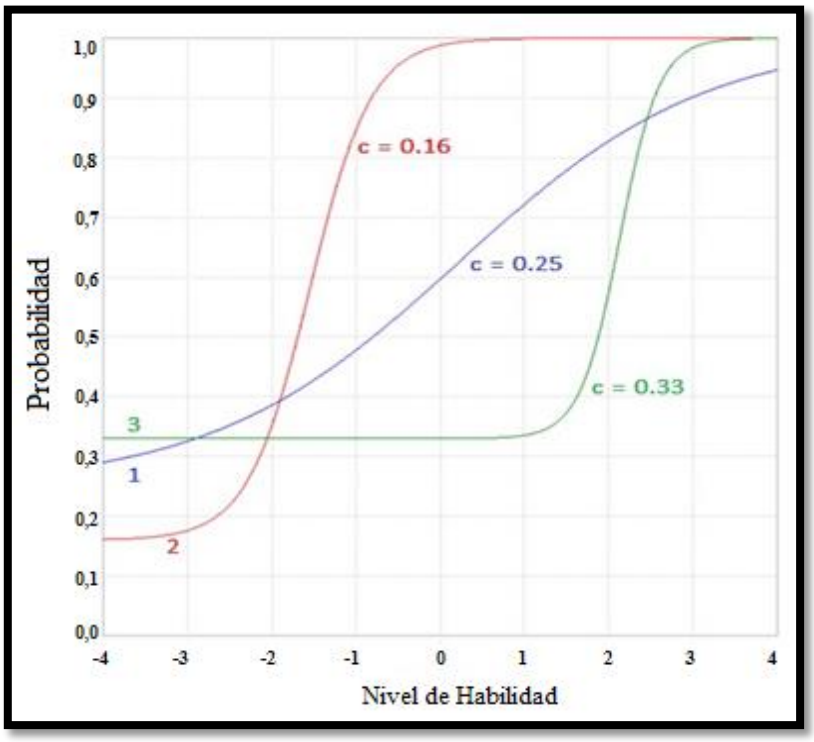


Figura 5. CCI Correspondiente a Preguntas con Diferentes Parámetros de Pseudo – Azar. *Curso de UML Multiplataforma Adaptativo Basado en la Teoría de Respuesta al ítem*. *Revista Ingeniería Informática, 10.*

Al interpretar la Figura 5 se puede observar que el parámetro de pseudo azar se asocia con el intercepto en el eje Y [3]; así la curva número 3 es la que contiene el parámetro de pseudo azar más alto lo que indica que este ítem es fácil de adivinar, contrario a lo que sucede con la curva número 2 la cual representa a un ítem más difícil de responder adivinando.

---

[3] La demostración se obtiene usando la ecuación característica [4]. Tenemos que el valor de $P(\theta_j)$ cuando θ_j tiende a $-\infty$, así,

$$P(\theta_j) = \lim_{\theta_j \to -\infty} c_i + (1 - c_i)\frac{e^{a_i(\theta_j - b_i)}}{1 + e^{a_i(\theta_j - b_i)}} \Rightarrow P(\theta_j) = c_i + (1 - c_i) \lim_{\theta_j \to -\infty} \frac{e^{a_i(\theta_j - b_i)}}{1 + e^{a_i(\theta_j - b_i)}} \Rightarrow P(\theta_j) = c_i + (1 - c_i)\frac{\lim_{\theta_j \to -\infty} e^{a_i(\theta_j - b_i)}}{\lim_{\theta_j \to -\infty}(1 + e^{a_i(\theta_j - b_i)})} \Rightarrow P(\theta_j) = c_i + (1 - c_i)\frac{e^{a_i b_i} \lim_{\theta_j \to -\infty} e^{a_i \theta_j}}{1 + e^{a_i b_i} \lim_{\theta_j \to -\infty} e^{a_i \theta_j}}$$

⇨ Por reglas de límites tenemos que $\lim_{x \to -\infty} e^x = 0$, así se da que $\Rightarrow P(\theta_j) = c_i + (1 - c_i)\frac{e^{a_i b_i \times 0}}{1 + (e^{a_i b_i \times 0})} \Rightarrow P(\theta_j) = c_i + ((1 - c_i) \times 0) \Rightarrow P(\theta_j) = c_i \blacksquare$

2.2.2.4. Supuestos del Modelo de TRI. La TRI según Muñiz (1997) se fundamenta en: "conseguir medidas invariantes respecto de los individuos medidos y de los instrumentos utilizados" (p.17), la invarianza en los parámetros de los ítems significa que los parámetros de la pregunta no cambian aunque las personas que contesten sean distintas y la invarianza del nivel de habilidad respecto al instrumento utilizado significa que el rasgo latente del individuo no depende del test.

La clave para que esta teoría sea efectiva está en la unión de los conceptos de separación de parámetros e invarianza de los mismos.

Para conseguir este propósito dicha teoría plantea modelos matemáticos que asumen que la probabilidad de que una persona dé una determinada respuesta puede ser descrita a partir de la posición de la persona en el nivel de rasgo de aptitud latente (θ) y de los valores de los parámetros del ítem como ya se ha indicado. Por consiguiente, los dos supuestos que subyacen a esta teoría están en relación con la naturaleza del rasgo que se desea medir (supuesto de unidimensionalidad) y con las relaciones entre las respuestas a los ítems (independencia local).

Unidimensionalidad. Un ítem se considera unidimensional si las diferencias sistemáticas dentro de la varianza del ítem se deben solo a una fuente de varianza, es decir, una variable latente. Para validar este supuesto se necesita mostrar que un solo rasgo explica las respuestas a los ítems. Los modelos expuestos en la sección 4.2.2.1 se asumen unidimensionales porque suponen que la respuesta dada a un ítem depende del nivel que tiene las personas en un único rasgo (θ).

Como guía para este apartado consideramos a Ferrando (1996) quien inicia por sugerir que para tomar evidencia de la dimensionalidad de un test se utilice un análisis factorial, el fundamento del método es considerar que una puntuación observada en un ítem puede descomponerse en dos

componentes independientes: uno que es explicado por la influencia de uno o varios factores comunes y una fuente residual denominada error, su finalidad es reproducir aquella parte de la puntuación en el ítem que se explica por la presencia del factor común que en términos de varianza implica reproducir la varianza común y no de la residual.

Finalmente, en cuanto a criterios de análisis para determinar la dimensionalidad con un AF, el autor recomienda usar el coeficiente phi (coeficiente de correlación de datos en un análisis factorial) si el interés es sólo determinar el número de factores aunque no descarta la posibilidad de interpretar otros recursos como son:

- Pesos o cargas factoriales: deberían ser cero o muy pequeños en los demás factores diferentes al primero, se recomienda su uso si los ítems tienen índices de dificultad muy extremos (desde 0.1 hasta 0.9).
- % de varianza explicado: comparar la del primer factor respecto a la varianza común (no la total), un porcentaje superior al 40% podría ser un referente de unidimensionalidad (Carmines y Zeller, 1979) aunque esto dependerá de cuanto la varianza total de los ítems corresponde a varianza del error, ésta podría ser mucha dejando muy poco a la varianza común pero que podría ser explicada por un solo factor.
- Análisis de las correlaciones o covarianzas residuales en busca de una distribución uniforme en torno a cero y sin valores extremos.
- Análisis de residuales: deben tender a cero si hay presencia de unidimensionalidad.

Adicionalmente puede usarse el criterio de información de Akaike (AIC) para determinar si al añadir parámetros libres al modelo (más factores) se consigue un mejor ajuste, el modelo que da

más información será aquel con menor discrepancia pero con menos parámetros libres. Su implementación práctica es simple, seleccionar el modelo con el menor valor en este criterio.

Independencia Local. Esta condición implica que cada ítem es independiente de los otros e incorrelacionado con los demás ítems, después de controlar por otros factores (genero, edad, entre otros). En consecuencia, la probabilidad condicional de observar un patrón de respuesta dado un valor de rasgo latente particular es igual al producto de las probabilidades condicionales de los ítems.

De no cumplirse este supuesto se habla de dependencia local la cual no puede ser probada directamente porque el rasgo latente es no observado y no puede ser aislada de otros supuestos del modelo por cuanto todos los supuestos son probados simultáneamente al usar un estadístico global como el Chi cuadrado de Pearson para datos discretos multivariados o un estadístico de razón de verosimilitud, es decir que al validarse la unidimensionalidad también se tiene la independencia local.

Entre las posibles fuentes que generan dependencia local se destacan:

Dimensionalidad: El test debe hallar una dimensión, es decir, el test busca el nivel de habilidad del individuo en un tema específico por tanto se debe comprobar su unidimensionalidad siendo el análisis factorial el más utilizado.

Multidimensionalidad: si una dimensión está determinada por subdimensiones, los ítems que pertenecen a esa subdimensión podrían estar más relacionados entre sí que con ítems de otra subdimensión (Brandt, 2017).

Los testlets o paquetes de ítems: grupos de ítems que comparten un estímulo (enunciado) en común, suelen usarse para economizar tiempo pero atentan contra el supuesto de independencia

porque la probabilidad de que una persona responda correctamente un ítem del paquete puede ser mayor que la probabilidad de acierto ante un estímulo diferente que ya ha sido contestado incorrectamente. Esta violación impacta la estimación produciendo sesgo en la estimación del parámetro de dificultad, sobreestimación en la estimación de varianza y de la confiabilidad (Liu & Maydeu-Olivares, 2012).

***2.2.2.5. Interpretación de puntuaciones*.** En la TCT encontramos que no es clara la relación entre el rasgo latente medido y el resultado observado en el test, así se obtuviera con exactitud la puntuación verdadera, lo cual no permite realizar conclusiones respecto a un constructo con base al resultado dado. En cambio, en la TRI por medio de la estimación de parámetros que se dan según el modelo utilizado, se encuentra una relación directa con la dimensión que se pretende medir, esto permite una mejor interpretación de los resultados.

2.2.2.6. Error estándar de medición. La TCT y la TRI tienen otra característica que las diferencia basada en el error estándar de medición. En la primera, este error es una constante para cualquier nivel de la puntuación verdadera y en la otra permite que el error estándar varíe con base a la habilidad latente, en esta última verifica que la precisión relacionada con una medición es menor en los extremos lo que implica que no es constante en toda la escala. Esto conlleva a que una medición puede ser menos confiable si se presenta una falta de correspondencia ente la dificultad global del test (ya sea fácil o difícil) y el nivel que se desea medir de la persona. Por lo mencionado anteriormente, el error estándar en la Teoría de Respuesta al Ítem (TRI) está asociado a la dificultad de cada ítem y a la habilidad de cada individuo.

En un contexto práctico, los softwares especializados en TRI aportan un recurso llamado Función de información, que será descrito con detalle más adelante, y es a partir de esta función que en forma recíproca se analiza el error estándar en la estimación [7] ya que:

$$SE(\theta) = \frac{1}{\sqrt{I(\theta)}} \qquad [7]$$

2.2.2.7. Invarianza de los parámetros. La diferencia fundamental de la TRI con respecto a la TCT radica en la invarianza de los parámetros, que solo es característica de la TRI. La invarianza de los parámetros significa que, si los supuestos de un modelo TRI se cumplen para un grupo de personas y de ítems, se cumplen las siguientes características:

- Las propiedades de los ítems, es decir sus parámetros de dificultad y discriminación, no cambian al considerarlos o aplicarlos en diferentes muestras de personas; las propiedades serían las mismas en una muestra de personas dotadas y una muestra de personas menos capaces. En la Teoría Clásica de los Test (TCT) esto no es el caso: los índices asociados con una prueba generalmente son distintos en diferentes muestras de personas.
- Las medidas de habilidad de las personas, son los mismos independientemente de la muestra de ítems que se incluyan en la prueba.

A nivel práctico, Olea y Posonda (2002), presentan dos formas distintas de comprobar las invarianzas mencionadas anteriormente, en primer lugar para la invarianza de las estimaciones de los parámetros de los ítems, se puede dar calibrando el banco de ítems en dos submuestras, por ejemplo una puede ser formada por individuos de menor nivel de habilidad y la otra por individuos con alto nivel de habilidad, y si se utiliza el modelo 1P en estas submuestras se debería obtener

que la correlación entre los valores estimados b_i debe ser cercana a 1. Y finalmente en la invarianza de las estimaciones de θ_j se puede obtener por medio de una correlación entre los niveles de habilidad de los individuos que conforman la muestra con dos submuestras distintas de ítems, por ejemplo se clasificaría el test en dos subtest y se escogería para una submuestra los ítems fáciles del primer subtest y la otra sería conformada por los ítems difíciles del segundo subtest, y la correlación entre ambas estimaciones debería ser próxima a 1.

2.3. Calidad del Test

2.3.1. Confiabilidad del instrumento. La confiabilidad se refiere a la consistencia de una medida. Para evaluar la confiabilidad o consistencia se utiliza el alfa de Cronbach, el cual es más propio de la Teoría Clásica de los Test (TCT) que de la TRI, cuyo cálculo correlaciona la puntuación de cada ítem del instrumento con la puntuación total observada de cada individuo y luego se hace la comparación con la varianza de todas las puntuaciones de cada ítem, como se refleja en la ecuación [8]:

$$\alpha = \left(\frac{k}{k-1}\right)\left(1 - \frac{\sum_{i=1}^{k} \sigma_{y_i}^2}{\sigma_x^2}\right) \quad [8]$$

Donde

k: es el número de ítems en el instrumento

$\sigma_{y_i}^2$*: la varianza asociada con el ítem i*

σ_x^2*: la varianza asociada con las puntuaciones totales observadas*

El alfa de Cronbach tiene ciertos criterios para su medición, es recomendado para su interpretación que se tome de referencia valores de 0.65 a 0.8; los coeficientes de valores menores

a 0.5 son inaceptables especialmente para escalas unidimensionales. Adicionalmente se referencias otro elemento a evaluar en relación con la confiabilidad del instrumento:

Estabilidad del instrumento. La estabilidad tiene que ver con la posible disminución de la confiabilidad debido al paso del tiempo, en qué grado se obtienen las mismas medidas al aplicar dos veces el mismo instrumento mediando entre ambas tomas un tiempo determinado, así debería aplicarse la misma prueba al mismo grupo de individuos en dos momentos diferentes y luego correlacionar estos puntajes.

2.3.2. Validez. Dado que la validez tiene como propósito evaluar si el instrumento está midiendo realmente el atributo que desea medir, propósito que puede resultar complejo, no obstante existen diferentes herramientas propuestas para orientar esta tarea, expuestas por Streefkerk (2019):

- La validez interna mide el grado en el que se puede estar seguro de que una relación de causa y efecto establecida en un estudio no puede explicarse por otros factores. En este mismo sentido, se hace referencia a la validez facial la cual permite que un grupo de personas opine sobre el grado de claridad en el vocabulario, y comprensión de cada ítem, se puede tomar en cuenta la opinión de personas que hayan tomado el test, potenciales usuarios de los resultados del test o en general personas que tengan algún interés en los resultados del test. Cuando está evaluación la realiza un grupo de expertos, el proceso sube de nivel y se habla de una validez de contenido ya que ellos pueden opinar sobre la relevancia y representatividad de cada ítem del cuestionario, además de la revisión de aspectos como terminología, diseño, formato y estilo, etc.

- La validez facial es considerada positiva sólo si hay un nivel razonable de acuerdo entre el grupo de calificadores.
- La validez externa se refiere al grado en que los resultados de un estudio se pueden generalizar a otras situaciones, o grupos. Tiene que ver con la muestra de participantes, ésta puede ser mejorada por fijar experimentos en una forma más natural y usando muestreo aleatorio para seleccionar los participantes.

2.3.3. Función de Información del Test (FI). Además de la CCI, para cada ítem se tiene esta función que permite analizar el error cometido al medir el rasgo de interés, el valor que asuma indicará para qué niveles del nivel de habilidad ($\boldsymbol{\theta}$) el ítem aporta mediciones más precisas. Sobre esta función, Posonda (2002) resalta los siguientes hechos a nivel teórico:

- En virtud de la estimación por máxima verosimilitud se tiene que la distribución del estimado de $\boldsymbol{\theta}$ es normal, con media $\boldsymbol{\theta}$ y varianza igual a como se expresa en la ecuación [9].

$$\sigma^2_{(\hat{\theta}|\theta)} = \frac{1}{\sum_{i=1}^{n} \frac{P'(\theta)_i^2}{P_i Q_i}} \quad [9]$$

  Donde $P'(\theta)$ es la derivada del modelo TRI que se esté implementando, P_i probabilidad de acertar el ítem i dado un nivel de habilidad θ y Q_i= 1- P_i.

- Al sacar raíz cuadrada a la anterior expresión se obtiene el error típico de medida y como se ve a continuación [10] el denominador corresponde a lo que se llama información del test.

$$S_e = \frac{1}{\sqrt{I(\theta)}} \quad [10]$$

A continuación el autor describe la función de información para cada uno de los tres modelos logísticos. Es claro que la información de un test dependerá de los parámetros de discriminación de los ítems (mayores valores en el parámetro a, mayor valor de la información), parámetro de pseudo - azar (menores valores en el parámetro c, mayor información), número de ítems (a mayor longitud, mayor información) y la convergencia entre el nivel de rasgo $\boldsymbol{\theta}$ y los parámetros b de los ítems (cuanto más próximos sean, mayor información).

- Para el modelo 1P:

$$I(\theta) = D^2 \sum P_i Q_i \quad [11]$$

- Para el modelo 2P:

$$I(\theta) = D^2 \sum a^2 P_i Q_i \quad [12]$$

- Para el modelo 3P:

$$I(\theta) = D^2 \sum \frac{a^2 Q_i (P_i - c)^2}{P_i (1-c)^2} \quad [13]$$

De la ecuación [10], los factores que influyen en los errores estándar se identifican mediante las características de la función de información del test. Tshering (2016) presenta los siguientes factores de los que depende el error estándar.

- La calidad de los ítems del test: los ítems de mayor discriminación que indican que no se puede adivinar para acertar la respuesta correcta, están relacionados con los errores estándar pequeños.
- La correspondencia entre el nivel de habilidad del individuo y la dificultad del ítem: los test elaborados por ítems con parámetros de dificultad iguales al nivel de habilidad del individuo, se asocian con errores estándar pequeños.

- El número de ítems del test: el error estándar será menor a medida que la longitud del test sea mayor.

Finalmente resaltar que la representación gráfica a continuación (Figura 6) muestra que es viable analizar la FI para cada ítem y otra para el test la cual corresponde a la suma de las FI de los ítems que lo conforman. Para la FI del ejemplo vemos que el test resulta más informativo para valores centrales del rasgo, y así mismo que la FI de un ítem como el FI5 será mayor para niveles de habilidad altos y menor para niveles bajos, con lo cual el ítem será más útil para medir a los individuos más hábiles.

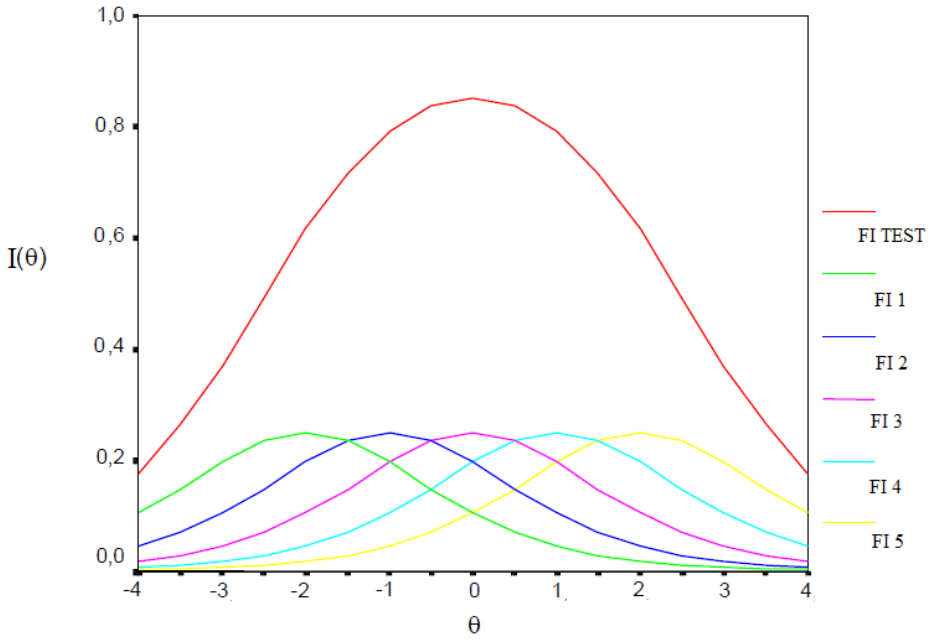


Figura 6. Funciones de Información de los ítems y del test. Adaptada de Olea y Ponsoda (2002). *Test Adaptativos Informatizados.* Madrid.

3. Metodología

El trabajo de grado realizado se enmarcó dentro de una metodología aplicada ya que lo que se hizo fue aplicar un modelo a un conjunto de datos provenientes de la aplicación de un test asociado

a temas de estadística descriptiva a un grupo de estudiantes. Los resultados se obtuvieron utilizando el software R.

Etapas:

1. Diseño del test: hace referencia a la selección y formulación de los ítems. Los ítems fueron extraídos de tests propuestos por DelMas, R., Garfield, J., y Ooms, A. (2005), Garfield, DelMas, Chance y Ooms (2006), Contreras, Cañadas, Gea y Arteaga (2013), GRE (2014), ICFES, Ziegler (2014), y algunos de creación propia. La selección se realizó incluyendo como dominios a evaluar: lectura e interpretación de gráficos (cuatro ítems), elementos teóricos básicos (cuatro ítems), interpretación de medidas de tendencia central (tres ítems), lectura e interpretación de tablas (cuatro ítems), interpretación de medidas de dispersión (seis ítems) y medidas de asociación (seis ítems). El formato de respuesta de los ítems incluidos en el test es dicotómico ya que para cada ítem hay una respuesta correcta pues lo que interesa es distinguir la capacidad de la persona de acertar la respuesta correcta.
2. La calibración del test incluyó las siguientes fases:

- Fase de pilotaje: se realiza un pilotaje inicial de las preguntas incluidas en el banco de ítems (ver Apéndice A), esta fase se implementó a través de una prueba de lápiz y papel aplicada a 52 estudiantes de los programas de matemáticas, licenciatura en matemáticas e ingeniería civil de la Universidad Industrial de Santander, quienes estaban cursando Estadística II y Estadística Aplicada a la Ingeniería en el semestre 2019-I.
- Fase de ajuste de los ítems: se realizaron los ajustes según los niveles de dificultad observados, validación de contenido y aspectos técnicos como longitud de la prueba y tiempo de duración principalmente.

- Fase de implementación del test: hechos los ajustes según la fase anterior, la prueba fue presentada por 66 estudiantes de los programas de matemáticas, licenciatura en matemáticas e ingeniería civil de la Universidad Industrial de Santander, quienes estaban cursando Estadística II y Estadística Aplicada a la Ingeniería en los semestres 2019-II y 2020-I. Con estos datos se ajustaron y evaluaron en los modelos logísticos de la TRI de un parámetro (1P), dos parámetros (2P) y tres parámetros (3P).

4. Resultados

Este capítulo inicia describiendo el análisis de resultados de la prueba piloto los cuales corresponden a la calibración del test. Posteriormente se presenta el ajuste hecho a los ítems y cómo se llegó a la versión final del test que fue el instrumento aplicado para tomar los datos que sirvieron de base para la fase final en la que se ajustaron y evaluaron los modelos de respuesta logísticos de un parámetro (1P), de dos parámetros (2P) y tres parámetros (3P).

4.1. Calidad del Test

Es claro que un test es válido si logra medir lo que se propone medir. En este sentido, la psicometría ofrece técnicas específicas para evaluar la calidad del test con base en dos características: la confiabilidad y la validez.

4.1.1. Confiabilidad. Al aplicar un instrumento se es consciente que al igual que toda medición ésta es susceptible a estar acompañada de errores, así es indispensable evaluar la precisión del

instrumento y cuál es su grado de consistencia interna para medir el constructo en cuestión. Es importante considerar que la confiabilidad se relaciona sólo con los errores aleatorios ya que puede haber presencia de errores sistemáticos que son explicados por otros factores o tienen una fuente identificable. Técnicamente un análisis de confiabilidad debería considerar tres aspectos: Consistencia interna, Estabilidad del instrumento y equivalencia.

La base para medir la confiabilidad son las correlaciones entre las preguntas del instrumento ya que cada una de estas evalúa el mismo atributo (constructo) de diferentes formas. El recurso usual para medir la consistencia interna es el coeficiente alfa de Cronbach cuyo valor para nuestro test es 0.65; dicha magnitud según los criterios de interpretación usual se ubicaría en el rango mínimo de aceptación pero dada la extensión del test que se propone, podríamos admitir que éste si logra evidenciar el constructo subyacente en cuanto al análisis de datos entre los participantes evaluados.

En nuestro caso no se abordó la evaluación de la estabilidad y equivalencia del test principalmente por motivos de tiempo en el primer caso y por no tener el número suficiente de tamaño de muestra para probar dos versiones diferentes de test tal como se requiere en el segundo caso.

4.1.2. Validez. Las preguntas que sirvieron de base para consolidar el test que se propone fueron obtenidas de las siguientes fuentes: test Comprehensive Assessment of Outcomes (CAOS), sitio web de ARTIST, test BLIS, GRE, ICFES, examen de estadística descriptiva de la Universidad de Granada y tres de creación propia; para las primeras tres fuentes mencionadas a continuación se resumen la evaluación de la validez del test que implementaron los autores (ver Tabla 1) y para efecto de nuestro análisis asumimos la validez de contenido que se hizo para los diferentes ítems ya que de nuestra parte sólo se realizó la traducción al español.

Tabla 1.

Evaluación de la validez de los test CAOS, ARTIST y BLISS.

Test	N° de participantes	Propósito	Análisis de confiabilidad (Consistencia)	Validez de contenido
CAOS	23.645	Evaluar razonamiento estadístico después del primer curso de Estadística (cultura estadística y entendimiento conceptual) con énfasis en el razonamiento en variabilidad.	Alfa de Cronbach = 0.78 Aceptable consistencia interna	Evaluación de un grupo de 18 expertos. La conclusión es que el test es una medida válida de importantes resultados de aprendizaje en un primer curso de Estadística.
ARTIST	555	Desarrollar un instrumento de evaluación práctico y accesible que permitan evaluar la alfabetización, razonamiento y pensamiento en un primer curso de estadística por parte de los estudiantes de cursos avanzados de estadística en secundaría y cursos introductorios en universidades.		Organizado por profesores miembros de instituciones de educación superior de USA que tenían experiencia que tenían experiencia en áreas afines de la estadística. La conclusión es que el test es válido para el estudio del razonamiento estadístico y es viable para comparar sus datos de referencia con futuras investigaciones.
BLISS	940	Evaluar métodos basados en simulación para cursos de introducción a la estadística a nivel postsecundario.	El coeficiente de alfa de Cronbach es 0.83	Supervisado por seis expertos en educación estadística. La conclusión es que el test proporciona información importante para estudios futuros sobre los conocimientos estadísticos en cursos introductorios de estadística que aborden temas de simulación.

Nota. Esta Tabla se hizo con base a la información establecida por los autores DelMas, R., Garfield, J., y Ooms, A. (2005 2006) y Ziegler (2014).

4.2. Desarrollo del Test

El tema central del test es la estadística descriptiva, los ítems están dirigidos a evaluar conocimientos y capacidad interpretativa de los datos y resultados estadísticos de los individuos.

Se consideraron seis dominios conceptuales: lectura de gráficos, uso de elementos teóricos, interpretación de medidas de tendencia central, lectura de tablas, interpretación de medidas de dispersión y medidas de asociación.

Para la prueba piloto, se seleccionaron 41 ítems de los cuales treinta y nueve eran de selección múltiple con única respuesta y dos ítems de respuesta abierta que fueron seleccionados de las fuentes mencionadas en la metodología, cada ítem tuvo una identificación única (nombre) durante todo el proyecto (ver Apéndice A), los ítems de respuesta abierta tenían como propósito explorar las posibles respuestas para convertirlos posteriormente en preguntas de selección múltiple.

A continuación se presenta el número de ítems de cada uno de los dominios mencionados anteriormente:

Tabla 2.

Clasificación de los ítems según los dominios conceptuales a evaluar.

Dominios a evaluar	ítems
Lectura e interpretación de Gráficos	P25,P19, P20,P22,P30,P37,P38,P41,P42
Elementos teóricos básicos	P1,P18,P24,P21,P17,P28,P31,P40
Interpretación de Medidas de Tendencia Central	P5,P6,P9,P35
Lectura e interpretación de Tablas	P8,P10, P36,
Interpretación Medidas de dispersión	P2,P3, P4, P23, P15,P16,P33,P34
Medidas de asociación	P7,P11, P12, P13,P14,P21,P29,P32,P39,P43

Debido a que era una cantidad considerable de ítems, no se podían aplicar todos en un solo test puesto que hubiera sido una prueba extensa para los estudiantes, por ello se decidió organizar los ítems en tres subconjuntos con ítems en común y otros diferentes, cada subconjunto tenía ítems de las seis competencias de aprendizaje, y esto condujo a la construcción de Test 1, Test 2 y Test 3, cada uno compuesto por 21 preguntas, cada uno con 14 ítems diferentes y 7 que son comunes en los tres test.

4.3. Aplicación del Test

4.3.1. Prueba piloto. El Test 1, Test 2 y Test 3 se aplicaron a estudiantes universitarios de los cursos Estadística II y Estadística Aplicada a la Ingeniería correspondientes a las carreras de ingeniería civil, Licenciatura en Matemáticas y matemáticas, en total participaron 52 estudiantes. A continuación se describe en la Tabla 3 la conformación de cada uno de estos tests:

Tabla 3.

Conformación de los Test de la Prueba Piloto.

Instrumento	Ítems
Test1	P1, P2,P3,P4,P5,P6,P7,P8,P9,P10, P11, P12,P15,P17,P18,P19,P28,P29,P30,P31,P32
Test2	P4,P5,P7,P8,P13,P16,P17,P18,P20,P21,P22,P28,P31,P32,P33,P34,P35,P36,P37,P38,P39
Test3	P4,P5,P7,P8,P12,P14,P16,P17,P18,P20,P21,P22,P23,P24,P25,P31,P32,P35,P40,P41,P42,P43

Para la aplicación, sin tener en cuenta el desempeño que llevaban en la materia en ese momento se les distribuyó de forma aleatoria uno de los test, el tiempo de respuesta estuvo comprendido entre noventa minutos y ciento cinco minutos. Después de la aplicación de los Test 1, Test 2 y Test 3, se llevó a cabo el análisis de las respuestas a los diferentes ítems a través del software R. Debido al tamaño de muestra el modelo TRI utilizado fue el modelo logístico de un parámetro (1P) con el cual se asume que el rendimiento en un ítem depende solamente del nivel de habilidad del individuo y de un único parámetro, la dificultad del ítem.

Al ajustar los datos a este modelo tendremos en cuenta que el parámetro de dificultad asume valores entre menos infinito e infinito. Iniciando con el Test 1 este se aplicó a 10 individuos, al ajustar los datos obtenidos al modelo de un parámetro se evidencia que aunque el programa estima la capacidad discriminatoria de los ítems éste es un valor fijo de 0.84 constante para todos los reactivos, se obtienen diferentes niveles de dificultad para los ítems, aproximadamente el 47,62% de los ítems (10) tienen un parámetro de dificultad menor a 0 es decir que fueron preguntas fáciles

de responder correctamente, y un 19.04% de los ítems (4) tienen un índice de dificultad mayor a 1.8 lo que se interpreta como preguntas difíciles de responder (ver Tabla 4). Debe notarse que hay un problema con los ítems P1, P4, P29 y P10, estos exhiben niveles de dificultad fuera del rango normal y además son valores muy negativos lo que debe interpretarse como que resultaron ser preguntas demasiado fáciles de responder en los primeros tres casos e imposible de responder en el caso de P10.

Tabla 4.

Estimación de la dificultad por ítem en el Test 1- Prueba Piloto.

Ítem	Dificultad	Discriminación
P1	-29,21	0,84
P4	-29,21	0,84
P29	-29,21	0,84
P3	-2,93	0,84
P12	-2,93	0,84
P15	-2,93	0,84
P5	-1,88	0,84
P8	-1,16	0,84
P7	-0,56	0,84
P30	-0,56	0,84
P6	0	0,84
P18	0	0,84
P19	0	0,84
P28	0	0,84
P31	0	0,84
P11	0	0,84
P9	0,56	0,84
P2	1,88	0,84
P17	2,93	0,84
P32	2,93	0,84
P10	29,21	0,84

El anterior análisis se complementa con la lectura de las CCI presentes en la Figura 7 las cuales, para facilitar la lectura, se presentan discriminando por dominio conceptual involucrado.

Inicialmente llama la atención las gráficas para los ítems P1, P4, P29 y P10 las cuales no describen crecimientos sigmoidales acorde a lo esperado en el modelo logístico de un parámetro, el comportamiento constante evidenciado en la Figura 7 permite confirmar con más claridad lo afirmado en el anterior párrafo en relación con estos ítems porque vemos que para P1, P4, P29 la probabilidad de respuesta correcta es de 1 sin importar el nivel de habilidad del individuo, situación opuesta se da con P10 que al parecer es el ítem más difícil de este test. Encontramos otros ítems fáciles como P5, P8, P12 y P15, ya que estudiantes con bajo nivel de habilidad pudieron responderlas correctamente.

De otro lado, las CCI de los ítems P18, P28, P31 coinciden con lo cual tienen igual nivel de dificultad y esta ojiva indica que habría un 50% de individuos con bajo nivel de habilidad que podrían acertar estas preguntas pertenecientes al dominio de elementos teóricos prácticos. En ítems como P2, P9, P17 y P32 al observar sus CCI se tiene que son preguntas que demandan un nivel de habilidad mayor a cero, es decir, es decir que individuos poco hábiles no podrán acertar la respuesta, cabe aclarar que estos ítems pertenecen a dominios de evaluación diferente. También se debe resaltar que no se observa mayor dificultad al responder preguntas de un dominio en particular, en cada uno hay preguntan de diferente nivel de dificultad.

Finalmente, debemos recordar que el comportamiento del Test 1 se está evaluando a partir de pocos datos (n=10) por cuanto un análisis más confiable requiere de más tamaño de muestra. En cuanto a las estimaciones de los niveles de dificultad se resalta como positivo que estas se distribuyen a lo largo de la escala lo que garantiza discriminación a lo largo de todo el rango de habilidad (Dificultad mínima=-3; Dificultad máxima=3) exceptuando los ítems anómalos P1, P4, P29 y P10.

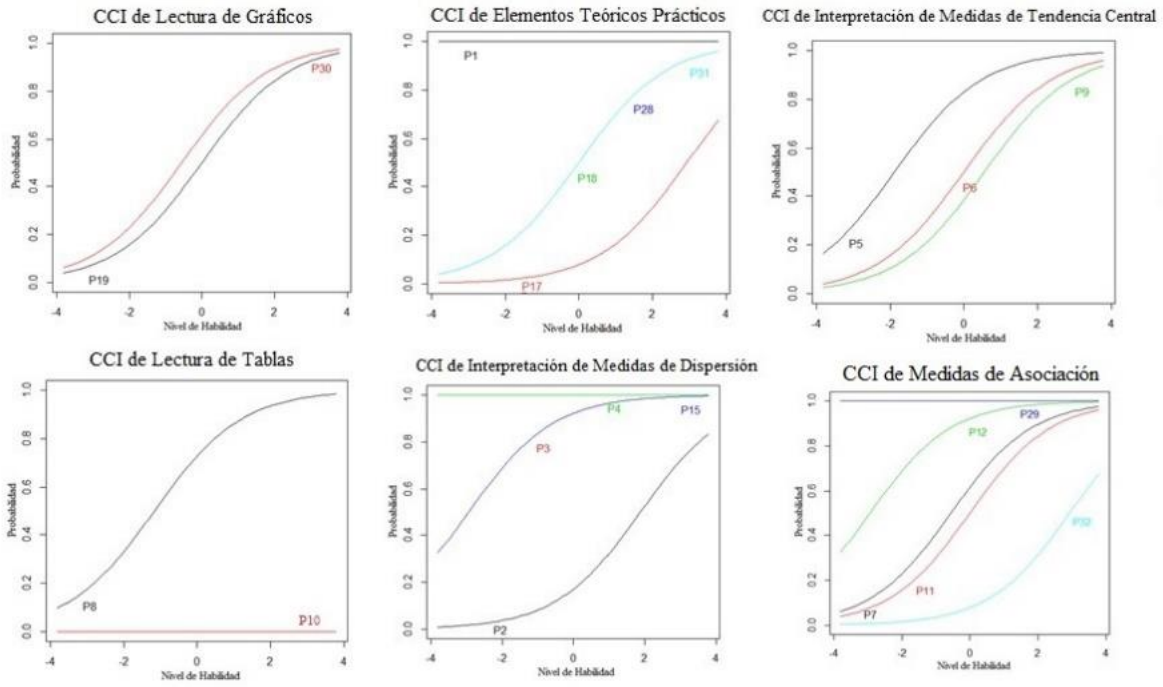


Figura 7. CCI de los ítems en el Test 1.

La Tabla 5 muestra en la tercera columna los valores p para la prueba de que los ítems se ajustan al modelo en análisis, todas resultan no ser significativas (α=0.05). No obstante, los ítems P1, P4, P29 y P10 presentaron una probabilidad de éxito de 1 y 0 para P10 lo que nos indica que al parecer no serán útiles en predecir el nivel de rasgo que se desea medir.

En general, para ítems con comportamiento inapropiado o pobre ajuste se podría pensar en eliminarlos pero esto reflejaría problemas con la validez de contenido de la prueba, por tanto en esta etapa del análisis se podría optar por la reformulación del ítem o volver a ponerlos en evaluación como fue nuestra decisión.

Tabla 5.

Ajuste de los ítems del Test 1 al modelo 1P – Prueba Piloto.

Ítem	X^2	Pr(>X^2)

P1	0	1
P2	8,54	0,29
P3	8,02	0,33
P4	0	1
P5	5,29	0,63
P6	9,07	0,25
P7	5,40	0,61
P8	8,46	0,29
P9	4,59	0,71
P10	0	1
P11	13,36	0,06
P12	6,55	0,48
P15	8,02	0,33
P17	8,84	0,26
P18	4,40	0,73
P19	4,21	0,76
P28	9,55	0,22
P29	0	1
P30	6,04	0,53
P31	4,21	0,76
P32	8,84	0,26

Continuando con el Test 2, éste se aplicó a 20 individuos, al ajustar los datos obtenidos al modelo de un parámetro se evidencia que el parámetro de dificultad no oscila en los rangos usuales ya que ítems como P4,P16, P33, P36, P37 y P39 tienen valores muy pequeños pero así mismo P13, P17 y P35 asumen valores muy grandes que como ya se dijo en el análisis al test 1 corresponden en ítems muy fáciles en el primer caso e ítems muy difíciles en el segundo caso; los niveles de dificultad observados se ubican entre -6 y 5, aproximadamente el 66,66% de los ítems (14) tienen un parámetro de dificultad menor a 0 es decir que la mayoría son preguntas fáciles de contestar correctamente, y sólo un 14.28% de los ítems (3) tienen un índice de dificultad mayor a 1.8. (Ver Tabla 6). En cuanto al parámetro de discriminación este es común para todos los ítems y se estimó en 0.46 valor que está por debajo del umbral mínimo (**a =0.65**).

Tabla 6.

Estimación de la dificultad por ítem para el Test 2- Prueba Piloto.

Ítem	Dificultad	Discriminación
P33	-6,57	0,46
P36	-6,57	0,46
P16	-6,57	0,46
P39	-3,90	0,46
P4	-3,13	0,46
P37	-3,13	0,46
P8	-2,48	0,46
P5	-1,92	0,46
P38	-1,40	0,46
P31	-1,40	0,46
P34	-0,92	0,46
P28	-0,92	0,46
P7	-0,92	0,46
P22	-0,45	0,46
P21	0,46	0,46
P32	0,92	0,46
P18	0,92	0,46
P20	0,92	0,46
P13	3,13	0,46
P17	3,90	0,46
P35	4,92	0,46

En las curvas características de los ítems en la Figura 8 se da que estas tienen poca pendiente debido a que el parámetro **a = 0.46** y conlleva a determinar que son ítems poco discriminativos; se evidencia de las CCI que los ítems P2, P13, P17 y P35 se tiene que si el nivel de habilidad se encuentra entre 0 a 4 se establecía una relación positiva ya que entre mayor sea el nivel de habilidad se tenía más probabilidad de responder acertadamente el ítem, y para estudiantes con niveles de habilidad menores a 0 era imposible que respondieran bien estos ítems. Las curvas características de los ítems P16, P33 y P36 nos indican que son ítems muy fáciles debido a que sin importar el nivel de habilidad se tiene una alta probabilidad de responderlos correctamente mayor a 0.8. Como apreciación general se tendría que decir que las CCI muestran que los ítems para evaluar medidas

de dispersión y lectura de tablas resultaron ser muy fáciles, en los otros dominios hay tanto ítems fáciles como difíciles.

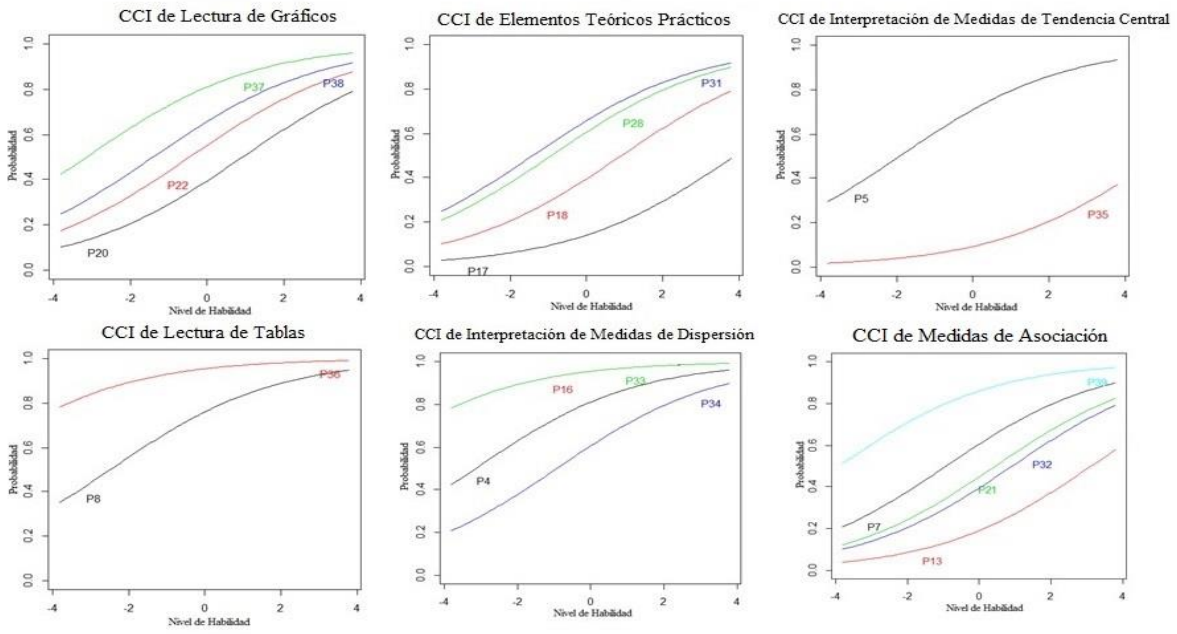


Figura 8. CCI de los ítems en el Test 2.

Finalmente, sobre el ajuste de los ítems del Test 2 al modelo de un parámetro (1P) sólo el ítem P7 no parece ajustar bien (ver Tabla 7) pero nuevamente se considera que el tamaño de muestra es bajo y se podría optar por volverlo a evaluar en la siguiente etapa del análisis.

Tabla 7.

Ajuste de los ítems del Test 2 al modelo 1P – Prueba Piloto.

	X^2	Pr(>X^2)
P4	3,76	0,71
P5	8,66	0,19
P7	17,48	0,0077
P8	3,47	0,75
P13	6,85	0,34
P16	4,98	0,55

P17	2,28	0,89
P18	3,31	0,77
P20	3,31	0,77
P21	2,80	0,83
P22	9,69	0,14
P28	0,68	0,99
P31	5,93	0,43
P32	3,56	0,74
P33	3,24	0,78
P34	4,52	0,61
P35	3,61	0,73
P36	4,98	0,55
P37	9,58	0,14
P38	8,11	0,23
P39	2,51	0,867

Para finalizar el análisis de la prueba piloto se examinarán los resultados para el Test 3 el cual fue aplicado a 22 individuos. Al ajustar los datos obtenidos al modelo de un parámetro se evidencia que el parámetro de dificultad, excepto para P4, P23, P17, P35 y P42, se ubica dentro de los límites sugeridos que son -4 y 4; se obtiene que el parámetro de discriminación para todos los ítems es aproximadamente de 0.31 el cual refleja que los ítems no cumplen con el umbral mínimo (**a =0.65**), Aproximadamente el 68,18% de los ítems (15) tienen un parámetro de dificultad menor a 0 es decir son preguntas fáciles de contestar correctamente, y solo un 13.63% de los ítems (3) tienen un índice de dificultad mayor a 1,8 por lo que serían calificadas como preguntas difíciles. (Ver Tabla 8).

Tabla 8.

Estimación de la dificultad por ítem para el Test 3- Prueba Piloto.

Ítem	Dificultad	Discriminación
P42	-82,99	0,31
P4	-7,59	0,31
P23	-6,10	0,31

P41	-4,98	0,31
P8	-4,98	0,31
P7	-2,53	0,31
P22	-1,86	0,31
P5	-1,86	0,31
P24	-1,86	0,31
P14	-1,86	0,31
P12	-1,22	0,31
P16	-1,22	0,31
P43	-1,22	0,31
P32	-0,61	0,31
P20	-0,61	0,31
P21	0	0,31
P25	0,61	0,31
P40	0,61	0,31
P31	1,22	0,31
P18	2,53	0,31
P17	7,59	0,31
P35	10,02	0,31

Se hace la aclaración de que la CCI de Lectura de Tablas se adicionó a la CCI de Interpretación de Medidas de Tendencia Central debido a que esta solo tenía un ítem (ver Figura 9). Como rasgo característico, las ojivas de los ítems se nota que tienen poca pendiente lo cual es un comportamiento coherente con el valor de **a** que es de 0.31, este es cercano a cero y conlleva a que los ítems sean poco discriminativos.

De las CCI de los ítems se observa que en su mayoría son ítem fáciles de responder, incluso P42 describe una probabilidad de 1 para todos los niveles de habilidad con lo cual resultó ser un reactivo demasiado fácil de responder acertadamente, esta misma situación se observa a P4, P8, P23 y P41 pero con una probabilidad un poco menor; por el contrario, P17 y P35 describen una probabilidad de éxito muy baja por lo que serían los ítems más difíciles de este test. Los demás ítems pueden considerarse que tienen una dificultad

El ítem 35 se podría considerar un ítem muy difícil ya que la CCI indica que estudiantes con niveles de habilidad mayor a 2 tendrían una probabilidad menor a 0.2 de responderlo acertadamente, el comportamiento de esta curva es debido a que hubo un error de redacción en la opción correcta de selección múltiple, lo que conllevaría a que los estudiantes consideraran las otras opciones cercanas a la respuesta. Debido a esto no se puede tomar la estimación del nivel de dificultad de este ítem. En cuanto a la dificultad evidenciada por dominios conceptuales, no se identifica ninguno en el que haya mayor nivel en ese parámetro.

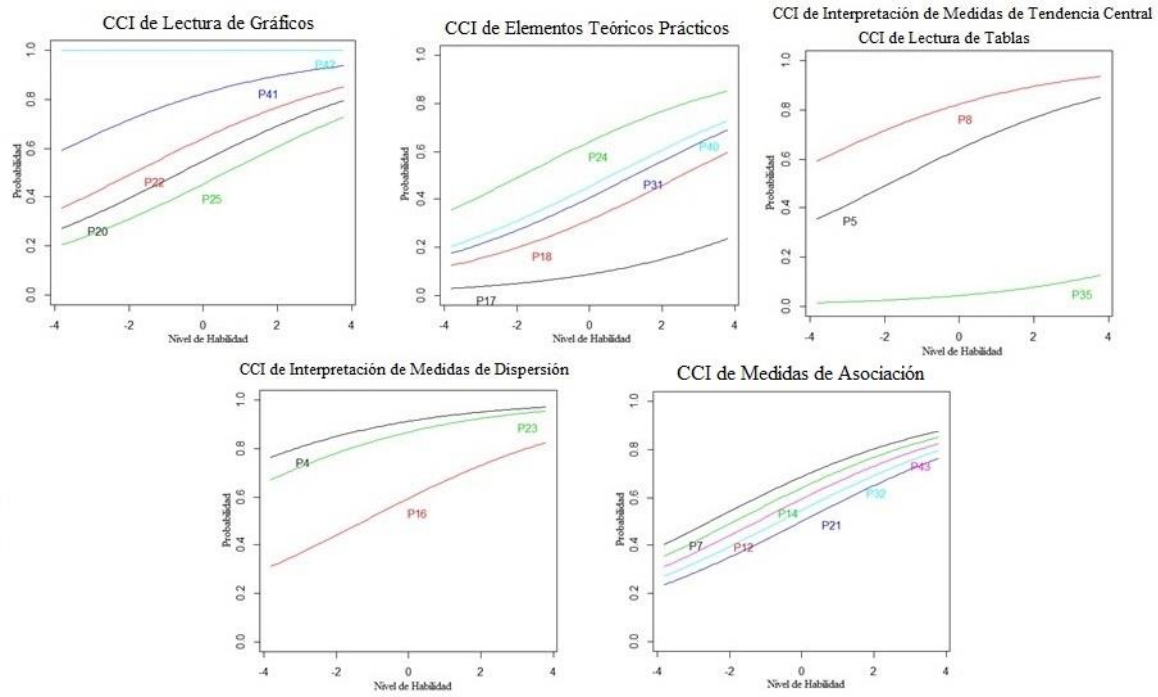


Figura 9. CCI de los ítems en el Test 3.

Finalmente sobre el análisis del ajuste de los ítems del Test 3 al modelo 1P, excepto por P5 para los demás se concluye buen ajuste (ver Tabla 9).

Tabla 9.

Ajuste de los ítems del Test 3 al modelo 1P- Prueba Piloto.

	X^2	Pr(>X^2)
P4	3,18	0,79
P5	16,49	0,01
P7	7,99	0,24
P8	4,43	0,62
P12	7,75	0,26
P14	7,24	0,30
P16	6,05	0,42
P17	3,43	0,75
P18	7,85	0,25
P20	3,24	0,78
P21	10,13	0,12
P22	2,11	0,90
P23	5,59	0,47
P24	4,68	0,59
P25	7,04	0,32
P31	5,52	0,48
P32	4,65	0,59
P35	5,64	0,46
P40	3,49	0,74
P41	4,26	0,64
P42	0	1
P43	7,37	0,29

4.3.2. Desempeño en la Prueba Piloto. Era importante realizar una Prueba Piloto debido a que se necesitaba además de hacer una estimación inicial de la dificultad de los ítems, seleccionar cuáles ítems eran adecuados para evaluar el razonamiento estadístico descriptivo acorde a un primer curso de Estadística en nuestro contexto académico, evaluar el comportamiento de las opciones de respuesta planteadas en cada pregunta para ir a ajustar el contenido del test definitivo. Como actividad secundaria se analizó el desarrollo de los dos ítems de respuesta abierta con el fin de tener insumos para la creación de preguntas propias que complementaran ciertos dominios de evaluación poco representados en el banco de ítems que se logró reunir (ver P7 y P21 en Apéndice A).

Para evaluar la eficacia de los distractores se analizó la información presente en la Tabla 10 dando prioridad a las preguntas que eran comunes a los tres test, la conclusión a que se llegó fue que las opciones incorrectas parecen estar bien planteadas ya que logran distribuir a los participantes en más de dos opciones con excepción de pocos ítems.

Tabla 10.

Porcentaje de estudiantes que eligieron cada opción de selección múltiple en los ítems de los de la prueba piloto.

Ítem	N° de Respuestas	a	b	c	d	e	N.R
P1	10	30*	0	0	70		
P2	10	20*	60	20	0		
P3	10	10	0	0	90*		
P4	52	1,92	1,92	88,46*	5,77		1,92
P5	52	1,92	69,23*	23,08	5,77		
P6	10	40	50*	10			
P7	52	26,92	63,46*	0	9,62		
P8	52	0	76,92*	5,77	17,31		
P9	10	40*	30	20	10		
P10	10	60*	30	0	10		
P11	10	0	30	20	50*		
P12	32	9,38	9,38	9,38	68,75*		3,13
P13	20	20*	10	5	65		
P14	22	0	0	63,64*	36,36		
P15	10	90*	0	0	10		
P16	42	7,14	11,9	76,19*	4,76		
P17	52	5,77	11,54	11,54*	1,92	69,23	
P18	52	34,62	38,46*	17,31	9,62		
P19	10	0	0	50*	10	40	
P20	42	4,76	2,38	7,14	38,1	47,62*	
P21	42	35,71	4,76	11,9	0	47,62*	
P22	42	2,38	0	19,05	59,52*	19,05	
P23	22	4,55	86,36*	0	9,09		
P24	22	4,55	13,64	18,18	0	63,64*	
P25	22	27,27*	40,91	18,18*	13,64		
P28	30	3,33	10	56,67*	30		
P29	10	30*	0	0	70*		
P30	10	10	20	60*	10		

P31	52	51,92*	19,23	5,77	23,08		
P32	52	53,85	40,38*	1,92	1,92		1,92
P33	20	0	5	95*	0		
P34	20	60*	25	10	5		
P35	42	23,81	19,05	7,14*	50		
P36	20	85*	0	10	5		
P37	20	80*	5	0	15		
P38	20	15	5	15	65*		
P39	20	75*	10	5	10		
P40	22	45,45*	22,73	13,64	18,18		
P41	22	4,55	81,82*	0	9,09		4,55
P42	22	100*	0	0			
P43	22	9,09	13,64	31,82	45,45*		

Nota. En esta tabla están todos los estudiantes que presentaron los tests ya que lo completaron en su totalidad. Los ítems sin resultados presentados para la opción D y E de selección múltiple representan un ítem que no tiene una opción D y E. * indica la respuesta correcta. *Nota* .Adaptada de Ziegler (2014).

Posterior a esto, se revisaron todos los ítems para resolver problemas de redacción y con base en las respuestas dadas a las preguntas abiertas se reformularon las preguntas P7 y P21 se plantearon dos nuevas preguntas relacionadas con lectura de tablas de contingencia, P26 y P27, tema que no había sido tenido en cuenta en el banco de ítems inicial.

Lo anterior aunado al análisis de las CCI de los 41 ítems, el cual fue presentado en la sección 4.3.1, sirvieron de base para definir la conformación del test a utilizar en la siguiente etapa al cual nos referiremos como Test Final y cuya conformación se presenta a continuación en la Tabla 11.

En el Apéndice B se incluyen las preguntas nuevas y las que fueron ajustadas. De esta forma, el Test Final quedó conformado por ítems que en cada dominio de evaluación son de diferentes niveles de dificultad y cada dominio incluye más de un ítem para su evaluación. El test final está construido con 27 ítems.

Tabla 11.

Clasificación de los ítems según la habilidad en el dominio conceptual requerido.

Dominios a evaluar	Ítems
Lectura e interpretación de Gráficos	P25,P19, P20,P22
Elementos teóricos básicos	P1,P18,P24,P17
Interpretación de Medidas de Tendencia Central	P5,P6,P9
Lectura e interpretación de Tablas	P8,P10,P26,P27
Interpretación Medidas de dispersión	P2,P3,P4, P23, P15,P16
Medidas de asociación	P7,P11, P12, P13,P14,P21

4.3.3. Administración del Test. El Test Final fue aplicado a estudiantes universitarios de los curso de Estadística II y Estadística Aplicada a la Ingeniería que se imparten para las carreras de Ingeniería Civil (2019-II) y Licenciatura en Matemáticas y Matemáticas en 2019-II y 2020-I.

En esta ocasión fueron evaluados 66 estudiantes diferentes a los que realizaron las pruebas piloto. El tiempo de aplicación del Test Final varío entre 40 minutos y 60 minutos, lo cual en comparación con la prueba piloto significó una reducción considerable.

4.4. Ajuste de los Modelos TRI

A continuación se presenta la implementación de las etapas usuales de un análisis basado en el ajuste de un modelo. El software utilizado fue R específicamente los paquetes stats, pysch, ltm y tpm. Se ajustaron los tres modelos logísticos de la TRI: 1 parámetro (1P), 2 parámetros (2P) y el modelo de 3 parámetros (3P) para luego realizar la comparación acorde a su bondad de ajuste.

En fase de ajuste de modelos se observó que los ítems P1, P4, P11 y P27 afectaban la calidad del ajuste debido a que son ítems detectados como fáciles de contestar correctamente, por consiguiente, fueron descartadas en el ajuste de datos.

4.4.1. Validación de supuestos. Para probar los dos supuesto básicos para un modelo TRI haremos uso de las correlaciones tetracóricas residuales para probar la unidimensionalidad

haciendo la comparación del cambio en el AF calculado a partir de la base de datos completa y la base de datos con testlets (grupos de ítems que comparten el mismo enunciado).

4.4.1.1. Unidimensionalidad e Independencia local. La implementación en R del análisis factorial se hizo con el comando factonal (paquete Stats de R) el cual permite contrastar la hipótesis de que un factor sea suficiente para describir la información presente en los datos; al usar todos los datos, es decir 27 variables y 66 individuos, la conclusión fue que un factor no es suficiente para resumir el comportamiento de los ítems del Test Final (Valor p= 0.03) lo que implica la violación al supuesto de unidimensionalidad.

Por lo anterior, se optó por considerar la presencia de testlets o grupos de ítems en el test, en nuestro caso se identificaron tres grupos que compartían el mismo enunciado: P12-P13-P14, P15-P16 y P26-P27. Se procedió a editar la base de datos para crear una variable dicotómica para cada grupo de ítems, el criterio fue que sólo si todas las respuestas a los ítems del grupo eran correctas, la nueva variable valdría 1, en caso contrario valdría cero. Al correr el análisis factorial con la base de datos ajustada, con 66 individuos y 23 ítems, se pudo descartar la necesidad de un modelo multidimensionalidad para describir estos datos (Valor P= 0.12) resultado que de paso nos garantiza la independencia local según se explicó en el marco teórico. A manera de complemento se presenta a continuación la matriz de correlaciones tetracóricas para la base de datos que incluye los testlets (ver Tabla 12), ésta fue obtenida con el comando tetrachoric del paquete psych de R.

Tabla 12.

Matriz de correlaciones tetracóricas.

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11
P1	1										
P2	-0,39	1									
P3	-0,04	0.39	1								

P4	0,02	0,2	0,31	1							
P5	-0,19	0,35	0,1	0,17	1						
P6	0,19	0,44	0,32	-0,08	0,3	1					
P7	-0,19	0,42	0,5	-0,12	0,12	0,11	1				
P8	0,21	0,36	0,64	-0,01	0,11	0,51	0,4	1			
P9	-0,22	0,27	-0,07	0,08	0,18	0,14	0,19	0,23	1		
P10	0,11	0,14	0,3	0,01	0,42	0,3	-0,11	0,29	-0,13	1	
P11	0,31	-0,05	0,09	0,35	-0,25	-0,05	0,11	0,18	0,04	0,05	1
P12-13-14	-0,03	0,3	-0,21	0,29	0,53	-0,05	0,24	-0,1	0,31	-0,04	0,39
P15-P16	-0,26	0,33	0,12	-0,14	-0,15	0,07	0,15	0,16	0,17	0,16	0,19
P17	-0,21	-0,17	0,16	-0,28	-0,14	0,04	0,19	0,25	0,13	-0,13	-0,24
P18	-0,19	0,17	0,39	0,24	0,26	0,13	0,07	0,36	-0,04	0,22	0,01
P19	-0,08	0,19	0,04	-0,06	0,16	0,17	0,21	0,22	0,32	-0,21	-0,24
P20	0,06	0,23	-0,02	0,07	-0,23	-0,02	0,19	0,01	0,25	-0,14	0,01
P21	0,03	0,08	0,38	0,08	-0,01	0,27	0,35	0,3	0,04	-0,34	0
P22	0,27	0,21	0,22	0,16	0,23	0,3	0,24	0,29	0,1	0,15	-0,1
P23	-0,1	0,45	0,59	0,26	0,3	0,41	0,24	0,61	0,49	0,12	-0,12
P24	-0,16	-0,02	0,24	0,11	-0,17	0,17	0,2	0,17	-0,01	0,06	0,17
P25	-0,49	0,09	0,01	0,25	0,3	-0,23	0,18	-0,38	0,34	-0,14	-0,32
P26-P27	-0,22	0,3	0,06	0,26	0,27	0,13	0,25	-0,02	-0,18	0,12	0,06

	P12-13-14	P15-P16	P17	P18	P19	P20	P21	P22	P23	P24	P25
P12-13-14	1										
P15-P16	-0,25	1									
P17	-0,32	-0,13	1								
P18	-0,11	0,11	-0,05	1							
P19	0,1	0,08	0,21	0,22	1						
P20	0,08	0,26	-0,09	0,01	0,56	1					
P21	-0,03	-0,06	0,2	0,25	0,54	0,07	1				
P22	0	0,05	0,08	-0,37	-0,15	0,02	-0,26	1			
P23	-0,12	0,26	0,03	0,33	0,24	0,02	0,1	0,43	1		
P24	-0,19	0,09	0,1	-0,17	-0,2	0,06	-0,03	0,3	0,08	1	
P25	0,17	-0,26	0,2	0	0,02	-0,11	0,15	-0,1	0	-0,22	1
P26-P27	0,14	0,07	-0,09	0,32	0,03	-0,01	0,26	0,01	-0,19	0,09	0,29

4.4.2. Modelo de Respuesta Logístico de un Parámetro (1P).

4.4.2.1. Coeficientes – Estimación de Parámetros. Al ajustar los datos obtenidos al modelo de un parámetro se evidencia que el parámetro de dificultad asume diferentes valores entre -3 y 3.

Para este modelo el índice de discriminación es común para todos los ítems y se ubica en 0,68 cumpliendo con el umbral mínimo (**a =0.65**). Aproximadamente el 60% de los ítems (13) tienen un índice de dificultad menor a cero lo que implica que son fáciles, así mismo, solo dos ítems tienen un parámetro de dificultad mayor a 1,8: P2 y P21 son las preguntas más difíciles que hacen parte de los dominios de interpretación de medidas de dispersión y medidas de asociación. (Ver Tabla 13).

Tabla 13.

Estimación de los parámetros por ítem del Test final-modelo 1P.

Ítem	Dificultad	Discriminación
P8	-2,95	0,68
P15	-2,75	0,68
P3	-2,41	0,68
P26	-2,41	0,68
P23	-2,11	0,68
P10	-1,84	0,68
P5	-1,59	0,68
P16	-1,47	0,68
P14	-1,24	0,68
P22	-0,91	0,68
P12	-0,60	0,68
P19	-0,60	0,68
P6	-0,30	0,68
P7	0	0,68
P18	0,10	0,68
P24	0,19	0,68
P25	0,39	0,68
P20	0,49	0,68
P17	0,49	0,68
P9	0,49	0,68
P13	1,12	0,68
P2	1,84	0,68
P21	2,58	0,68

Al graficar los valores de dificultad como se ilustra en la Figura 10 se puede observar que en los ítems P9 y P13 hay un hueco grande que se da entre 0,49 y 1,12, lo que nos indicaría que se deben construir ítems con nivel de dificultad entre estos dos valores, para elaborarlos se debe estudiar con detalle las características de ambos ítems lo cual se haría en una posible continuidad de este proyecto, este caso se presenta también entre los ítems P13 y P2, P2 y P21.

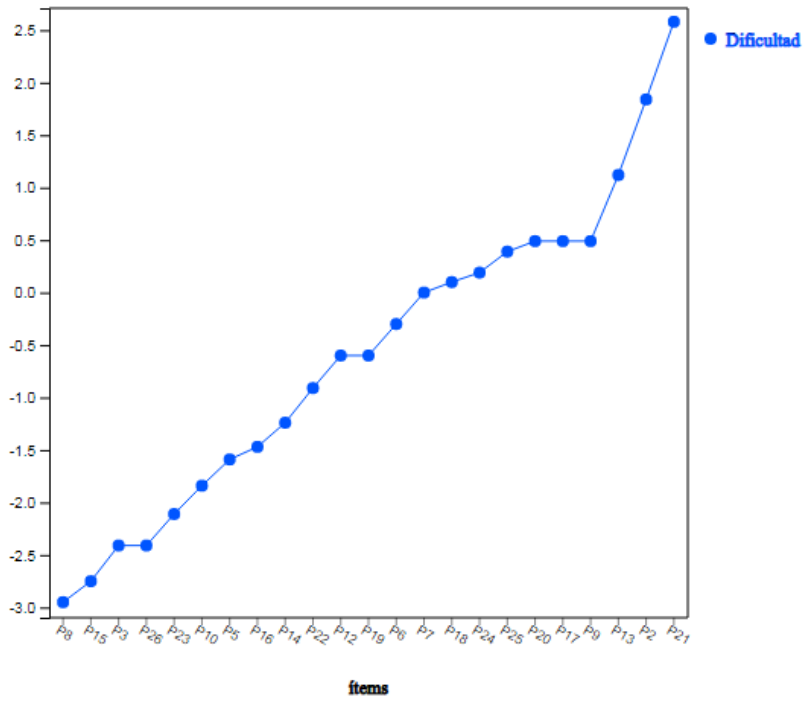


Figura 10. Parámetros de Dificultad de los ítems en el modelo 1P.

4.4.2.2. Puntuación en el Factor. La Tabla 14 es producida con el comando factor.scores (paquete ltm de R), presenta para cada patrón de respuesta el número observado de estudiantes que lo tuvieron (Obs) y el número esperado de ocurrencias (Exp), bajo la columna denominada z1 aparece el puntaje en el factor, es decir la habilidad estimada para cada patrón de respuesta y finalmente el error estándar para dicha estimación. Para nuestros datos se observa que no hubo dos estudiantes que hubiesen contestado con el mismo patrón de respuesta en este conjunto de 66 participantes (ver valores en columna Obs). Los niveles de habilidad son iguales en algunos

patrones de respuesta, por ejemplo, los individuos 11, 22, 27, 40, 45, 46, 48, 54 tienen el mismo patrón de respuesta contiene catorce unos y nueve ceros (ver Apéndice C), y poseen el mismo nivel de habilidad, 0.13, aunque poseen respuestas iguales sólo en dos ítems el P15 y P16 que fueron correctas.

Continuando el individuo 65 tiene un patrón de respuesta con 21 unos y dos ceros (ver Apéndice C); el individuo 66 le corresponde un patrón de respuestas con veintidós unos y un cero, aunque estos estudiantes poseen respuestas iguales en 22 ítems el nivel de habilidad de cada uno es distinto, la diferencia radicó en el ítem P21 (ítem con mayor nivel de dificultad) puesto que la persona que lo respondió correctamente obtuvo un nivel de habilidad de 2,09 y la persona que lo respondió incorrectamente un nivel de habilidad de 1,81.

Tabla 14.

Patrones de respuesta observados vs esperados y habilidad estimada para cada uno-1P.

	Obs	Exp	z1	se,z1		Obs	Exp	z1	se,z1
1	1	0	-1,38	0,57	34	1	0	0,57	0,58
2	1	0	-1,38	0,57	35	1	0	0,57	0,58
3	1	0	-0,95	0,56	36	1	0	-0,95	0,56
4	1	0	-0,95	0,56	37	1	0	-0,09	0,56
5	1	0	-1,60	0,57	38	1	0	-0,52	0,56
6	1	0	-1,38	0,57	39	1	0	0,35	0,57
7	1	0	-1,16	0,56	40	1	0	0,13	0,57
8	1	0	-1,16	0,56	41	1	0	-0,09	0,56
9	1	0	-0,09	0,56	42	1	0	0,35	0,57
10	1	0	0,57	0,58	43	1	0	-0,09	0,56
11	1	0	0,13	0,56	44	1	0	-0,09	0,56
12	1	0	-0,73	0,56	45	1	0	0,13	0,57
13	1	0	-1,16	0,56	46	1	0	0,13	0,57
14	1	0	-0,73	0,56	47	1	0	0,57	0,58
15	1	0	-1,16	0,56	48	1	0	0,13	0,57
16	1	0	-0,52	0,56	49	1	0	0,80	0,59
17	1	0	-0,30	0,56	50	1	0	1,04	0,59
18	1	0	-0,30	0,56	51	1	0	0,80	0,59

19	1	0	-1,16	0,56	52	1	0	0,35	0,57
20	1	0	-0,52	0,56	53	1	0	0,80	0,59
21	1	0	0,35	0,57	54	1	0	0,13	0,57
22	1	0	0,13	0,57	55	1	0	-0,09	0,56
23	1	0	-0,52	0,56	56	1	0	1,04	0,59
24	1	0	-0,52	0,56	57	1	0	-0,30	0,56
25	1	0	-0,73	0,56	58	1	0	0,57	0,58
26	1	0	-0,73	0,56	59	1	0	-0,30	0,56
27	1	0	0,13	0,57	60	1	0	1,04	0,59
28	1	0	0,35	0,57	61	1	0	0,80	0,59
29	1	0	0,35	0,57	62	1	0	0,80	0,59
30	1	0	-0,30	0,56	63	1	0	0,80	0,59
31	1	0	0,35	0,57	64	1	0	1,54	0,62
32	1	0	0,80	0,59	65	1	0,02	1,81	0,63
33	1	0	0,80	0,59	66	1	0,01	2,09	0,65

4.4.2.3. Precisión (CCI). En la Figura 11 podemos observar que las curvas características de los ítems tienen poca pendiente, esto es debido a que el parámetro **a** estimado es 0,68 lo cual sugiere poca discriminación. En cuanto a la interpretación del nivel de dificultad (**b**) a partir de las CCI hay varios aspectos por indicar. Un primer hecho evidente en Figura 11 son las CCI para los ítems P2 y P21 ya que muestran que se requerirá un nivel de habilidad cercano a 2 para tener un chance del 40% de una respuesta correcta por lo cual, estos dos ítems son los más difíciles del test.

Continuando con el análisis de la Figura 11 se puede concluir que los ítems que evaluaron elementos teóricos prácticos se pueden denominar de nivel promedio ya que estudiantes con niveles de habilidad mayor a cero tienen una probabilidad superior al 50% de responderlos correctamente. También se debe notar que los ítems relacionados con lectura de tablas e interpretación de medidas de dispersión, excepto el ítem P2, indican que son preguntas fáciles ya que aún estudiantes con niveles de habilidad bajos como lo es -2 tienen probabilidades superiores al 40% de responderlos adecuadamente; finalmente resaltar que los ítems relacionados con

medidas de asociación y medidas de tendencia central fueron los que describieron mayor variación en sus niveles de dificultad (rango de variación 3,82 y 2,08 respectivamente).

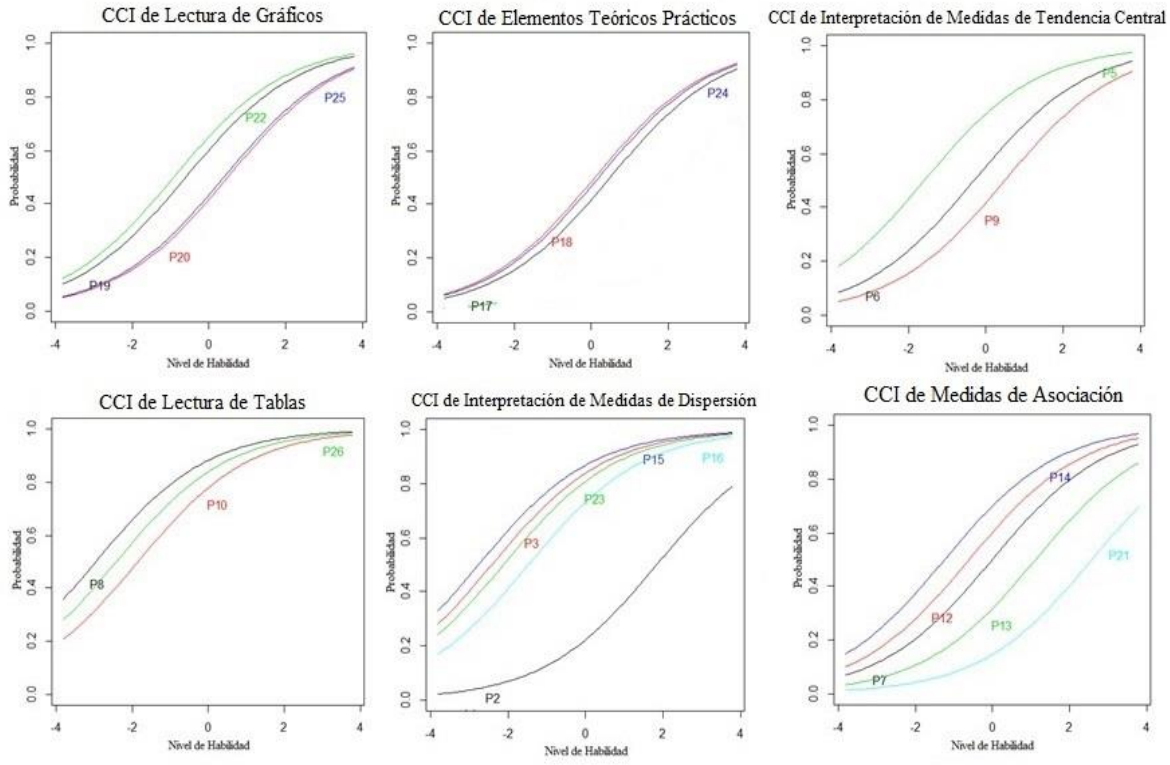


Figura 11. CCI de los ítems en el modelo 1P.

4.4.2.4. Función de Información del Test (FI). En la Figura 12 se muestra la función de información. Recuérdese que, su importancia radica en que permite evaluar la precisión de la estimación para calcular intervalos de confianza para las estimaciones de los niveles de habilidad. El comportamiento deseable es que la curva sea simétrica respecto al nivel de habilidad cero pero en este caso la FI del Test Final ajustado al modelo de un parámetro es asimétrica a derecha indicando que este instrumento proporciona mejor información, es decir menor error, para los estudiantes con niveles de habilidad bajos, entre -2 y 1.5 con lo cual hay poca precisión en la estimación de los niveles altos de habilidad, al comparar con los resultados de la Tabla 14 se

determina que el 95.45% de los individuos (63) obtuvieron niveles de habilidad entre los valores mencionados.

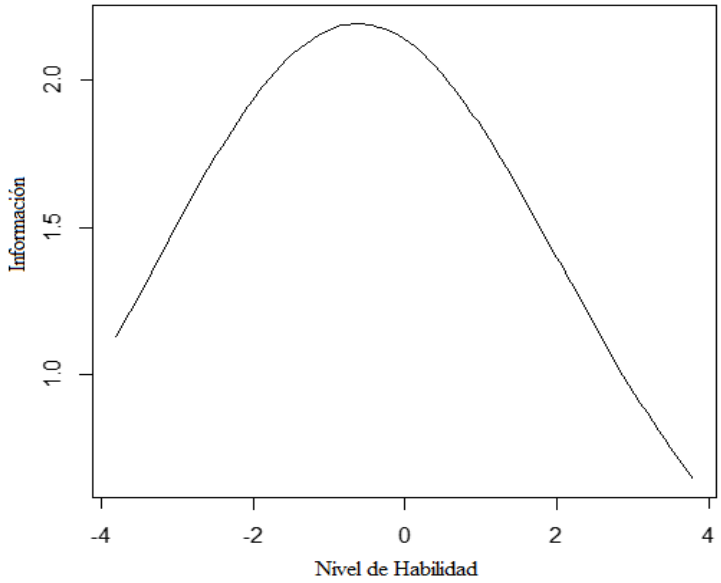


Figura 12. Función de Información del Test en el modelo 1P.

4.4.2.5. Ajuste de la persona. Esta salida se obtiene a través del comando person.fit de R el cual tiene por objeto aportar un medio para hacer detección de valores extremos a nivel de participantes. Toma como soporte el estadístico basado en la verosimilitud para determinar el patrón de respuesta más típico dado un modelo específico. El estadístico que se aporta para hacer la evaluación aparece denotado por Lz el cual según Paek y Cole (2020) se interpreta como una z-puntaje usual con límites para detección de ±3.0 de cada participante en relación con la estimación de la habilidad. En la Tabla 15 no se observan valores extremos, sólo en la fila 2 y fila 44 Lz < -2 pero que no se detectan como patrones de respuesta inesperados en relación con el modelo de un parámetro (1P).

Tabla 15.

Patrón de respuesta de la persona y su correspondiente habilidad estimada (Lz)-1P.

	L0	Lz	Pr(<Lz)		L0	Lz	Pr(<Lz)
1	-14,01	-0,34	0,37	34	-13,31	-0,26	0,40
2	-18,34	**-2,53**	0,01	35	-11,57	0,55	0,71
3	-13,05	0,33	0,63	36	-16,46	-1,47	0,07
4	-15,65	-1,05	0,15	37	-11,53	1,05	0,85
5	-14,26	-0,59	0,28	38	-12,59	0,62	0,73
6	-13,91	-0,29	0,38	39	-11,61	0,72	0,76
7	-15,47	-1,01	0,16	40	-10,81	1,28	0,90
8	-13,58	-0,02	0,49	41	-10,41	1,63	0,95
9	-11,71	0,96	0,83	42	-13,27	-0,09	0,46
10	-11,18	0,74	0,77	43	-11,02	1,32	0,91
11	-16,30	-1,49	0,07	44	-18,05	**-2,34**	0,01
12	-15,04	-0,69	0,24	45	-11,81	0,78	0,78
13	-16,45	-1,51	0,07	46	-10,84	1,27	0,90
14	-13,70	0,02	0,51	47	-10,76	0,93	0,82
15	-10,44	1,61	0,95	48	-16,46	-1,57	0,06
16	-16,95	-1,71	0,04	49	-8,71	1,64	0,95
17	-14,81	-0,60	0,27	50	-8,47	1,47	0,93
18	-13,25	0,23	0,59	51	-13,41	-0,48	0,31
19	-15,57	-1,05	0,15	52	-12,02	0,52	0,70
20	-14,86	-0,59	0,28	53	-12,22	0,06	0,52
21	-10,43	1,30	0,90	54	-12,77	0,29	0,62
22	-13,63	-0,14	0,44	55	-13,59	-0,02	0,49
23	-13,55	0,10	0,54	56	-12,15	-0,12	0,45
24	-13,25	0,27	0,60	57	-11,53	1,14	0,87
25	-9,94	2,03	0,98	58	-10,75	0,94	0,83
26	-9,40	2,32	0,99	59	-12,11	0,83	0,80
27	-11,69	0,84	0,80	60	-12,35	-0,21	0,42
28	-12,43	0,32	0,62	61	-10,59	0,79	0,79
29	-13,81	-0,35	0,36	62	-10,80	0,70	0,76
30	-12,38	0,69	0,76	63	-10,91	0,65	0,74
31	-9,96	1,53	0,94	64	-8,60	0,85	0,80
32	-14,03	-0,76	0,22	65	-6,29	1,45	0,93
33	-8,98	1,52	0,94	66	-6,17	1,18	0,88

4.4.2.6. Ajuste de los ítems. El ajuste de los ítems del Test Final al modelo de un parámetro, según los valores p en la Tabla 16 todos son no significativos excepto P25 por cuanto debería evaluarse su formulación o removerlo del test.

Tabla 16.

Ajuste de los ítems al modelo – 1P.

Ítem	X^2	Pr(>X^2)
P2	10,23	0,18
P3	9,75	0,20
P5	6,44	0,49
P6	6,83	0,45
P7	8,91	0,26
P8	9,70	0,21
P9	6,26	0,51
P10	5,94	0,55
P12	8,85	0,26
P13	2,74	0,91
P14	10,73	0,15
P15	7,06	0,42
P16	4,87	0,68
P17	7,34	0,39
P18	12,12	0,10
P19	5,64	0,58
P20	6,35	0,50
P21	10,04	0,19
P22	8,62	0,28
P23	8,98	0,25
P24	12,02	0,10
P25	18,44	0,01
P26	6,30	0,51

4.4.3. Modelo de Respuesta Logístico de dos Parámetro (2P).

4.4.3.1. Coeficientes – Estimación de Parámetros. El resultado del ajuste de los datos del Test Final al modelo logístico de dos parámetros (2P) aparece en la Tabla 17. Sobre el parámetro de

dificultad, las estimaciones obtenidas se ubican en un rango entre -4 y +4 lo cual es deseable porque hay diferentes niveles de dificultad en el test, como se ve a partir del ordenamiento de los ítems el más fácil fue P17 seguido de P16, P10 y P26 y como los ítems más difíciles se perfilan P13 y P21, los demás ítems podrían calificarse que tienen una dificultad entre moderada y promedio (**b** es cercano a 0).

De otro lado en cuanto al parámetro de discriminación recordemos que este es el encargado de informar que tan bien el ítem discrimina entre los diferentes niveles de habilidad ($\boldsymbol{\theta}$) y que su escala usual oscila entre 0 y 3. Acorde a los resultados en la Tabla 17 y asumiendo como criterios de interpretación la propuesta de Baker (2001), se observa que las estimaciones se ubican en el rango normal excepto tal vez por el ítem P17 que se muestra levemente negativo; el 41.67% de los ítems (10) tienen un índice de discriminación entre 0 a 0.61 lo cual indica que no alcanzan el umbral para tener un funcionamiento aceptable ($\boldsymbol{a_i \geq 0.65}$), no obstante el 17.69% de los ítems (5) tienen un elevado nivel de funcionamiento ($\boldsymbol{a_i > 1.34}$) destacándose entre ellos los ítems P23, P8 y P2 que son altamente discriminatorios ($\boldsymbol{a_i > 1.69}$) y abordan temas de dispersión y lectura de tablas.

Tabla 17.

Estimación de los parámetros por ítem del Test final-modelo 2P.

Ítem	Dificultad	Discriminación
P17	-3,95	-0,08
P16	-2,52	0,37
P10	-2,12	0,58
P26	-1,82	0,97
P15	-1,64	1,40
P3	-1,47	1,35
P8	-1,31	2,81
P5	-1,29	0,88
P22	-1,25	0,47

P23	-1,09	1,93
P14	-0,91	1,02
P19	-0,77	0,51
P12	-0,48	0,91
P6	-0,21	1,12
P7	0,00	1,00
P18	0,11	0,61
P9	0,50	0,66
P2	0,81	2,90
P24	1,15	0,11
P20	1,25	0,25
P25	1,76	0,14
P13	2,32	0,31
P21	3,47	0,49

En la Figura 13 se puede observar que con respecto a los parámetros de dificultad obtenidos en el modelo 2P, los ítems P17 y P16 presentan un hueco grande que se da entre -3,95 y -2,52, lo que nos proyectaría a que se deberían construir ítems con nivel de dificultad entre estos dos valores, este caso se presenta también entre los ítems P13 y P21, P20 y P25.

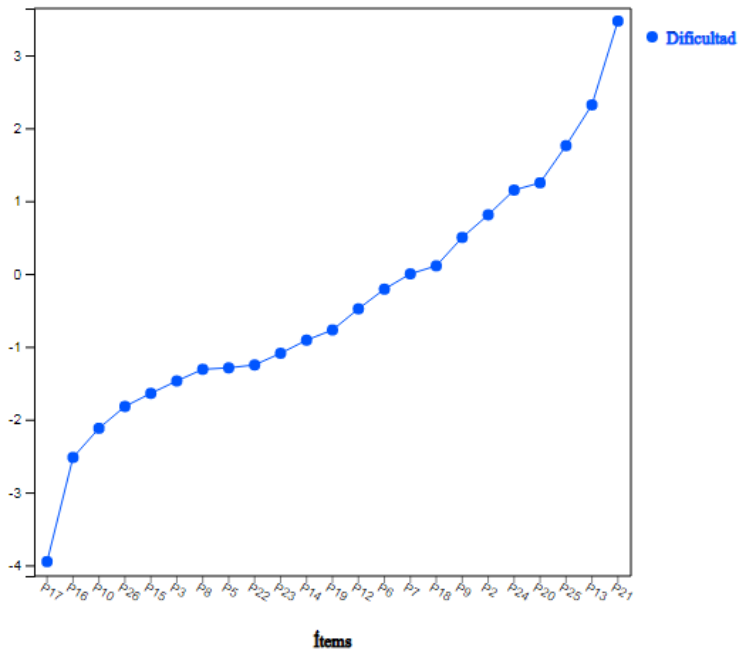


Figura 13. Parámetros de Dificultad de los ítems en el modelo 2P.

4.4.3.2. Puntuación en el Factor. Los datos ajustados al modelo de dos parámetros evidencian que no hay dos estudiantes que hubiesen contestado con el mismo patrón de respuesta por eso la segunda columna de la Tabla 18 es uno. Los niveles de habilidad son diferentes en los 66 patrones de respuesta, tomando un caso específico tenemos en los individuos 11, 22, 27, 40, 45, 46, 48, 54 sus niveles de habilidad varían en un rango de -0.69 a 0.75, el patrón de respuesta de estos participantes contiene catorce unos y nueve ceros (ver Apéndice C), aunque solo poseen respuestas iguales en dos ítems, el P15 y P16 que fueron correctas; en la mayoría de patrones de respuesta bajó considerablemente el nivel de habilidad comparado con los resultados dados en el modelo 1P puesto que en este para todos θ=0.13, exceptuando la línea 54 ya que el nivel de habilidad aumento.

Continuando con el análisis de patrones de respuesta, en la fila 65 el patrón tiene veintiún unos y dos ceros (ver Apéndice C) mientras que el de la fila 66 contiene veintidós unos y un cero, aunque estos estudiantes poseen respuestas iguales en 22 de los ítems el nivel de habilidad de cada uno es distinto, la diferencia radico en el ítem P21 puesto que la persona que lo respondió correctamente obtuvo un nivel de habilidad de 1,86 y la persona que lo respondió incorrectamente un nivel de habilidad de 1,66; en comparación con el nivel de habilidad dado en el modelo de un parámetro (Tabla 14) en el modelo 2P han disminuido considerablemente los niveles de habilidad de los patrones de respuestas.

Tabla 18.

Patrones de respuesta observados vs esperados y habilidad estimada para cada uno-2P.

	Obs	Exp	z1	se,z1		Obs	Exp	z1	se,z1
1	1	0	-1,53	0,40	34	1	0	0,13	0,48
2	1	0	-1,71	0,42	35	1	0	0,42	0,47
3	1	0	-1,10	0,40	36	1	0	-1,24	0,39
4	1	0	-0,78	0,42	37	1	0	-0,03	0,48

5	1	0	-1,64	0,42	38	1	0	-0,27	0,47
6	1	0	-1,66	0,42	39	1	0	0,27	0,48
7	1	0	-1,49	0,40	40	1	0	0,02	0,48
8	1	0	-1,05	0,40	41	1	0	0,18	0,48
9	1	0	0,11	0,48	42	1	0	-0,04	0,48
10	1	0	0,14	0,48	43	1	0	0,11	0,48
11	1	0	-0,69	0,43	44	1	0	-0,46	0,46
12	1	0	-1,00	0,40	45	1	0	0,15	0,48
13	1	0	-1,12	0,40	46	1	0	0,29	0,48
14	1	0	-0,92	0,41	47	1	0	0,45	0,47
15	1	0	-0,65	0,44	48	1	0	0,06	0,48
16	1	0	-0,84	0,42	49	1	0,01	0,61	0,46
17	1	0	-0,79	0,42	50	1	0	0,66	0,46
18	1	0	-0,31	0,47	51	1	0	1,07	0,49
19	1	0	-1,20	0,39	52	1	0	0,85	0,46
20	1	0	-0,42	0,46	53	1	0	1,13	0,50
21	1	0	0,25	0,48	54	1	0	0,75	0,46
22	1	0	0,17	0,48	55	1	0	0,18	0,48
23	1	0	-0,39	0,46	56	1	0	0,99	0,48
24	1	0	-0,89	0,41	57	1	0	0,68	0,46
25	1	0	-0,54	0,45	58	1	0	0,98	0,48
26	1	0	-0,46	0,46	59	1	0	0,48	0,47
27	1	0	0,15	0,48	60	1	0	1,19	0,51
28	1	0	0,01	0,48	61	1	0	1,10	0,49
29	1	0	0,34	0,47	62	1	0	1,27	0,53
30	1	0	-0,31	0,47	63	1	0	0,95	0,47
31	1	0	0,14	0,48	64	1	0,01	1,57	0,60
32	1	0	0,21	0,48	65	1	0,02	1,66	0,62
33	1	0	0,57	0,46	66	1	0,01	1,86	0,67

4.4.3.3. Precisión (CCI). Para la lectura de la CCI asociadas a un modelo 2P se analizan dos características gráficas: el desplazamiento a izquierda o derecha de la ojiva que permite identificar los ítems fáciles de los difíciles y la pendiente de la curva que está en relación con el nivel de discriminación.

En cuanto al nivel de dificultad se verifican que las ojivas se desplazan por los diferentes valores de la escala horizontal y describen diferentes formas, hacia el lado derecho se destacan como los ítems más difíciles P25, P13 Y P21, del otro extremo se identifican los ítems con tendencia a la facilidad siendo las más extremos P17, P16 y P10. La clasificación de ítems según el nivel de dificultad cambia considerablemente respecto de la que aporta el modelo 1P.

En cuanto al poder discriminatorio de los ítems, P2 (pregunta sobre medidas de dispersión) y P8 (pregunta de lectura de tablas) poseen la pendiente más alta en comparación a las demás curvas de los ítems lo que indica que estos ellos son los de mayor poder discriminatorio. También, es evidente que la CCI del ítem P17 presenta un comportamiento anómalo ya que su gráfica es casi lineal y decreciente (pendiente negativa) por lo cual registra un índice de discriminación en cero además de ser el ítem más fácil del test. Del ítem P8, P24 y P25 podemos decir que son medio discriminativos debido a que estudiantes con nivel de habilidad cercanos a -1 prácticamente no tienen posibilidad de responderlo acertadamente pero estudiantes con nivel de habilidad mayor a -1 tienen una alta probabilidad de responderlo correctamente.

En los seis dominios se puede evidenciar que las CCI de la mayoría de sus ítems (70%) tienen una pendiente baja lo que implica tener ítems poco discriminativos con lo cual estudiantes con poca habilidad tienen cierto nivel de probabilidad de responderlos correctamente. (Ver Figura 14).

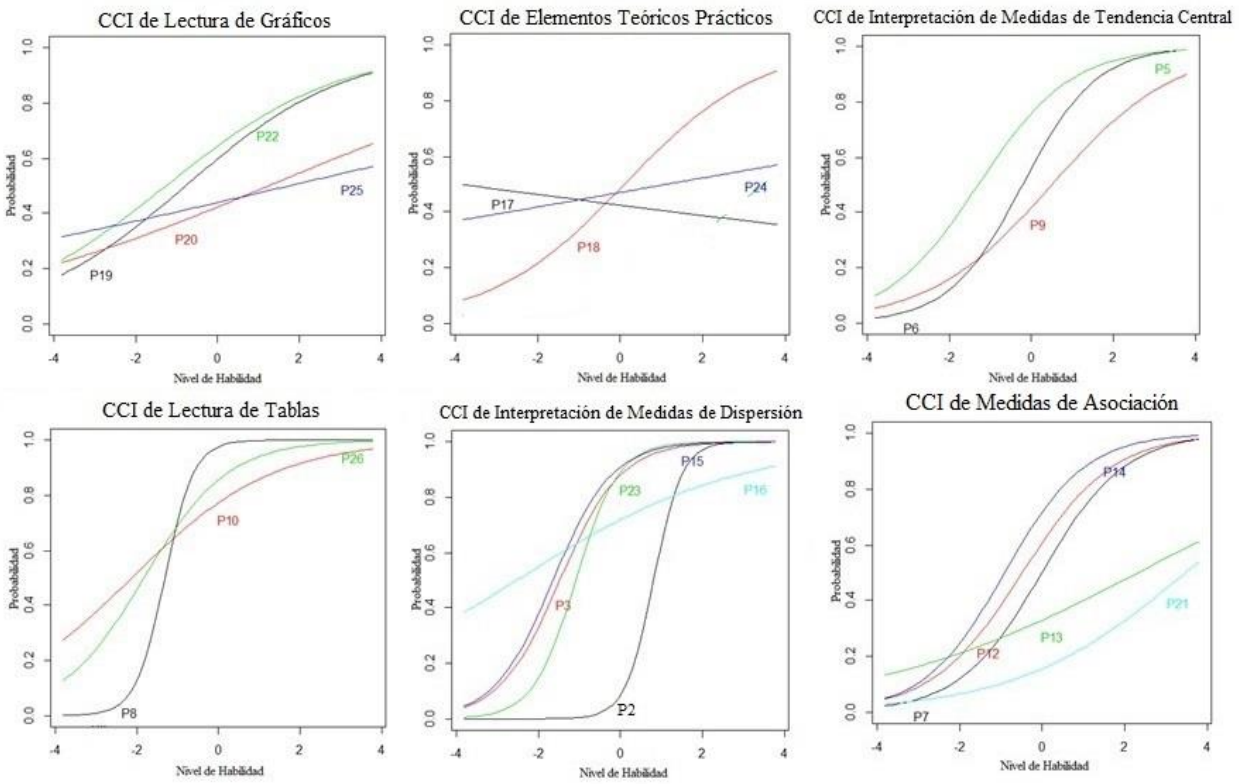


Figura 14. CCI de los ítems en el modelo 2P.

4.4.3.4. Función de Información del Test (FI). En la Figura 15 se evidencia que la función de información del test no posee un comportamiento simétrico pero se puede interpretar que el Test Final ajustado con el modelo de dos parámetros proporciona mayor información para los estudiantes con niveles de habilidad entre -2 y -1 y da una información considerable para los estudiantes con niveles de habilidad comprendidos de 0 a 1. Según esto la conclusión es que el modelo no es tan preciso al estimar las habilidades en los extremos superiores es decir tanto los menos hábiles como los más hábiles.

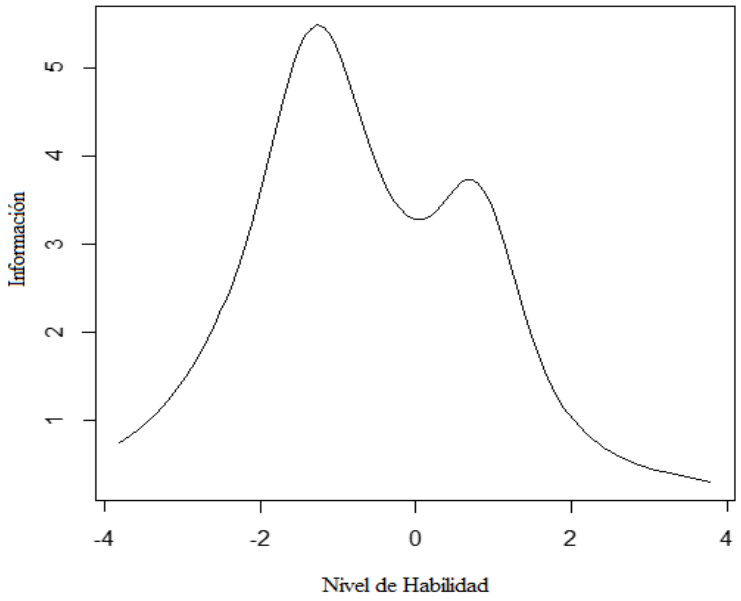


Figura 15. Función de Información del Test en el modelo 2P.

4.4.3.5. Ajuste de la persona. En la medida de ajuste de la persona en la fila 44 que se muestra en la Tabla 19 usando de referencia un valor pequeño Lz < -2 se detecta una cadena de respuestas problemático o aberrante potencial en relación con el modelo de dos parámetros (2P).

Tabla 19.

Patrón de respuesta de la persona y su correspondiente habilidad estimada (Lz) – 2P.

	L0	Lz	Pr(<Lz)		L0	Lz	Pr(<Lz)
1	-12,26	0,76	0,78	34	-13,76	-0,62	0,27
2	-14,36	-0,69	0,25	35	-11,19	0,43	0,67
3	-11,61	1,55	0,94	36	-15,43	-1,01	0,16
4	-15,67	-1,14	0,13	37	-11,01	0,86	0,80
5	-12,39	0,51	0,69	38	-12,66	0,22	0,59
6	-10,61	1,49	0,93	39	-11,24	0,51	0,70
7	-13,19	0,25	0,60	40	-10,44	1,09	0,86
8	-13,68	0,18	0,57	41	-9,84	1,26	0,90
9	-11,58	0,47	0,68	42	-13,09	-0,17	0,43
10	-11,64	0,41	0,66	43	-10,54	0,97	0,83

11	-17,05	-1,98	0,02	44	-17,50	**-2,26**	0,01
12	-13,41	0,35	0,64	45	-11,36	0,54	0,70
13	-15,30	-0,89	0,19	46	-10,15	1,03	0,85
14	-13,44	0,30	0,62	47	-10,26	0,85	0,80
15	-11,75	1,09	0,86	48	-16,43	-1,89	0,03
16	-15,96	-1,31	0,10	49	-8,48	1,59	0,94
17	-13,41	0,24	0,59	50	-8,90	1,35	0,91
18	-12,80	0,18	0,57	51	-12,47	-0,75	0,23
19	-14,21	-0,18	0,43	52	-10,91	0,20	0,58
20	-14,53	-0,66	0,25	53	-10,67	0,00	0,50
21	-10,11	1,08	0,86	54	-11,84	-0,14	0,44
22	-13,12	-0,33	0,37	55	-14,97	-1,24	0,11
23	-13,53	-0,14	0,44	56	-12,07	-0,49	0,31
24	-12,70	0,74	0,77	57	-11,05	0,30	0,62
25	-9,92	2,00	0,98	58	-9,31	0,80	0,79
26	-9,64	2,04	0,98	59	-12,45	-0,22	0,41
27	-11,14	0,65	0,74	60	-11,06	-0,25	0,40
28	-12,30	0,18	0,57	61	-9,12	0,74	0,77
29	-13,13	-0,45	0,33	62	-8,36	0,87	0,81
30	-12,28	0,46	0,68	63	-10,28	0,38	0,65
31	-9,74	1,34	0,91	64	-7,63	0,84	0,80
32	-14,75	-1,15	0,12	65	-6,35	1,29	0,90
33	-8,91	1,41	0,92	66	-6,82	0,89	0,81

4.4.3.6. Ajuste de los ítems. Los valores p de la siguiente Tabla 20 no son significativos excepto tal vez el de P25 con lo cual la conclusión es que los ítems tienen un buen ajuste al modelo 2P.

Tabla 20.

Ajuste de los ítems al modelo – 2P.

Ítem	X^2	Pr(>X^2)
P2	7,80	0,45
P3	7,29	0,51
P5	7,49	0,48
P6	7,71	0,46
P7	10,07	0,26
P8	2,16	0,98
P9	6,40	0,60
P10	8,30	0,40

P12	5,57	0,70
P13	7,89	0,44
P14	8,59	0,38
P15	3,08	0,93
P16	3,66	0,89
P17	2,33	0,97
P18	9,89	0,27
P19	7,26	0,51
P20	8,42	0,39
P21	9,96	0,27
P22	8,24	0,41
P23	5,03	0,75
P24	6,13	0,63
P25	16,13	0,04
P26	5,59	0,69

4.4.4. Modelo de Respuesta Logístico de tres Parámetros (3P).

4.4.4.1. Coeficientes – Estimación de Parámetros. Según la tabla a continuación**,** el índice de dificultad varía en un rango de -4 y +4, lo cual indica que hay variación en este parámetro, sólo se exceptúa el ítem P24 que presenta un elevado índice de dificultad de 11,98. En cambio, las estimaciones para el parámetro de discriminación no lucen muy bien a según el criterio de Baker (2001), se tienen aproximadamente un 26% de valores demasiado grandes para este parámetro, entre 28 y 231, información que se complementará más adelante con la observación de las CCI (Figura 17).

En cuanto al parámetro de pseudo-azar observamos que en 22 ítems oscila entre 0 a 0,5 que es el rango en el que normalmente se encuentra este índice (Olea y Ponsoda, 2002), recordar que este parámetro está en relación con el número de respuestas de cada pregunta, en general para los ítems en evaluación concluiríamos que sólo con P26 existiría la preocupación de que éste ítem pudiera llegar a ser fácil de adivinar (Ver Tabla 21)

Tabla 21.

Estimación de los parámetros por ítem del Test final- modelo 3P.

Ítem	Dificultad	Discriminación	Pseudo-azar
P16	-3,53	0,25	0,02
P10	-2,89	0,41	0
P22	-1,35	0,43	0
P3	-1,28	1,71	0
P23	-1,03	2,03	0
P8	-0,71	42,42	0,29
P15	-0,67	1,85	0,5
P6	-0,24	0,87	0
P7	0,01	1,2	0
P26	0,07	2,33	0,65
P18	0,1	0,66	0
P5	0,17	2,98	0,5
P14	0,24	230,82	0,48
P12	0,67	136,34	0,38
P25	0,69	63,84	0,3
P19	0,72	1,16	0,38
P9	0,79	1,88	0,2
P2	0,91	1,95	0
P17	1,41	28,82	0,4
P13	1,42	2,49	0,25
P20	1,82	58,62	0,41
P21	2,55	0,68	0
P24	11,98	0,02	0,06

En la Figura 16 se puede observar que los ítems P21 y P24 tienen parámetros de dificultad en el modelo 3P comprendidos entre 2,55 y 11,98, lo cual es un hueco grande y como ya se había mencionado en el modelo 1P y 2P se tendría que examinar la elaboración de ítems con nivel de dificultad entre estos dos valores, este caso se presenta también entre los ítems P16 y P10, P10 y P22, P20 y P21.

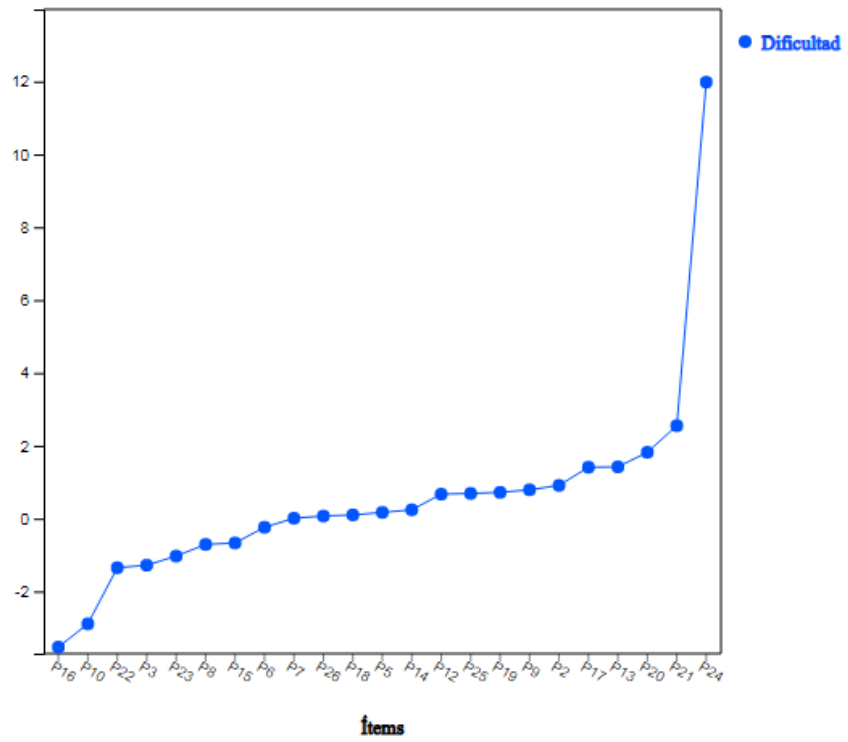


Figura 16. Parámetros de Dificultad de los ítems en el modelo 3P.

4.4.4.2. Puntuación del Factor. Los datos ajustados al modelo de tres parámetros evidencian que no hay dos estudiantes que hubiesen contestado con el mismo patrón de respuesta por eso la segunda columna de la Tabla 22 es uno. Los niveles de habilidad son diferentes en los 66 patrones de respuesta, tomando el mismo caso que hemos usando en los dos modelos estudiados anteriormente tenemos en los individuos 11, 22, 27, 40, 45, 46, 48, 54 el patrón de respuesta con catorce unos y nueve ceros (ver Apéndice C), aunque solo poseen respuestas iguales en dos ítems el P15 y P16 que fueron correctas, sus niveles de habilidad varían en un rango de -0,811 a 0,76. En la mayoría de patrones de respuesta aumentó el nivel de habilidad comparado con el asignado a cada uno en el modelo 2PL, exceptuando la fila 11, 22 y 48 ya que en estos patrones de respuesta el nivel de habilidad disminuyó.

Utilizando el segundo caso planteado en los otros dos modelos ya estudiados se obtuvo que en la fila 65 el patrón de respuesta con 21 unos y dos ceros; en la fila 66 correspondiente al patrón de respuestas con veintidós unos y un cero, aunque estos estudiantes poseen respuestas iguales en 22 ítems el nivel de habilidad de cada uno es distinto, la diferencia como ya se ha indicado radica en el ítem P21 puesto que el estudiante que lo respondió correctamente obtuvo un nivel de habilidad de 1,935 y el estudiante que lo respondió incorrectamente un nivel de habilidad de 1,891; en el modelo 2P han aumentado considerablemente los niveles de habilidad de los patrones de respuestas (Tabla 18) y en comparación con el nivel de habilidad dado por el modelo 1P (Tabla 14) con el modelo 3P hubo disminución.

Tabla 22.

Patrones de respuesta observados vs esperados y habilidad estimada para cada uno-3P.

	Obs	Exp	z1	se,z1		Obs	Exp	z1	se,z1
1	1	0	-1,54	0,58	34	1	0	0,26	0,10
2	1	0	-1,17	0,49	35	1	0	0,76	0,17
3	1	0	-1,50	0,60	36	1	0	-1,04	0,47
4	1	0	-1,01	0,52	37	1	0	0,26	0,07
5	1	0	-1,56	0,60	38	1	0	-0,18	0,46
6	1	0	-1,56	0,60	39	1	0	0,21	0,06
7	1	0	-1,72	0,56	40	1	0	0,15	0,47
8	1	0	-1,66	0,55	41	1	0	0,05	0,49
9	1	0	0,04	0,58	42	1	0	0,01	0,49
10	1	0	-0,04	0,61	43	1	0	0,32	0,48
11	1	0	-0,81	0,21	44	1	0	-0,12	0,46
12	1	0	-0,95	0,48	45	1	0	0,22	0,06
13	1	0	-0,63	0,14	46	1	0	0,49	0,44
14	1	0	-0,78	0,10	47	1	0	1,45	0,12
15	1	0	-0,30	0,44	48	1	0	0,27	0,20
16	1	0	-0,61	0,20	49	1	0,01	0,77	0,19
17	1	0	-0,59	0,29	50	1	0,01	0,76	0,17
18	1	0	-0,28	0,45	51	1	0	0,68	0,02
19	1	0	-1,15	0,49	52	1	0	0,27	0,13
20	1	0	-0,09	0,42	53	1	0	0,67	0,02

21	1	0	-0,06	0,42	54	1	0	0,76	0,17
22	1	0	0,17	0,43	55	1	0	0,21	0,08
23	1	0	-0,12	0,40	56	1	0	1,51	0,18
24	1	0	-0,77	0,10	57	1	0	0,36	0,44
25	1	0	-0,24	0,55	58	1	0	0,68	0,02
26	1	0	-0,08	0,54	59	1	0	0,22	0,05
27	1	0	0,27	0,15	60	1	0	0,90	0,43
28	1	0	-0,10	0,55	61	1	0	0,84	0,49
29	1	0	0,47	0,51	62	1	0	0,68	0,02
30	1	0	-0,12	0,52	63	1	0	0,65	0,04
31	1	0	0,37	0,45	64	1	0,01	1,21	0,36
32	1	0	0,73	0,08	65	1	0,25	1,89	0,16
33	1	0,01	0,76	0,16	66	1	0,19	1,94	0,42

4.4.4.3. Precisión (CCI). Para interpretar la información gráfica que se muestra en la Figura 17 empezaremos por analizar las CCI de todos los ítems en relación con el nivel de dificultad, los ítems más difíciles son P24, P20 y P21. Con relación a la discriminación, el rasgo a examinar es la pendiente, aquí se confirma lo dicho sobre el comportamiento anómalo para los ítems P8, P12, P14, P17 y P20 ya que el índice de discriminación para el ítem j ($\mathbf{a_j}$) es proporcional a la pendiente de la CCI en el valor $\boldsymbol{\theta} = b_j$ y vemos que en este caso dicha recta es prácticamente vertical. Finalmente, sobre el parámetro de pseudo-azar se observaría la asíntota inferior de la CCI (intercepto con el eje vertical) para identificar la probabilidad que tiene una persona de responder acertadamente la pregunta con un nivel de habilidad extremadamente bajo, de esta forma en el ajuste del modelo 3P, sólo llama la atención P26 que tiene asociada una probabilidad de más 0.6 para un individuo con una habilidad baja de $\boldsymbol{\theta} = -4$.

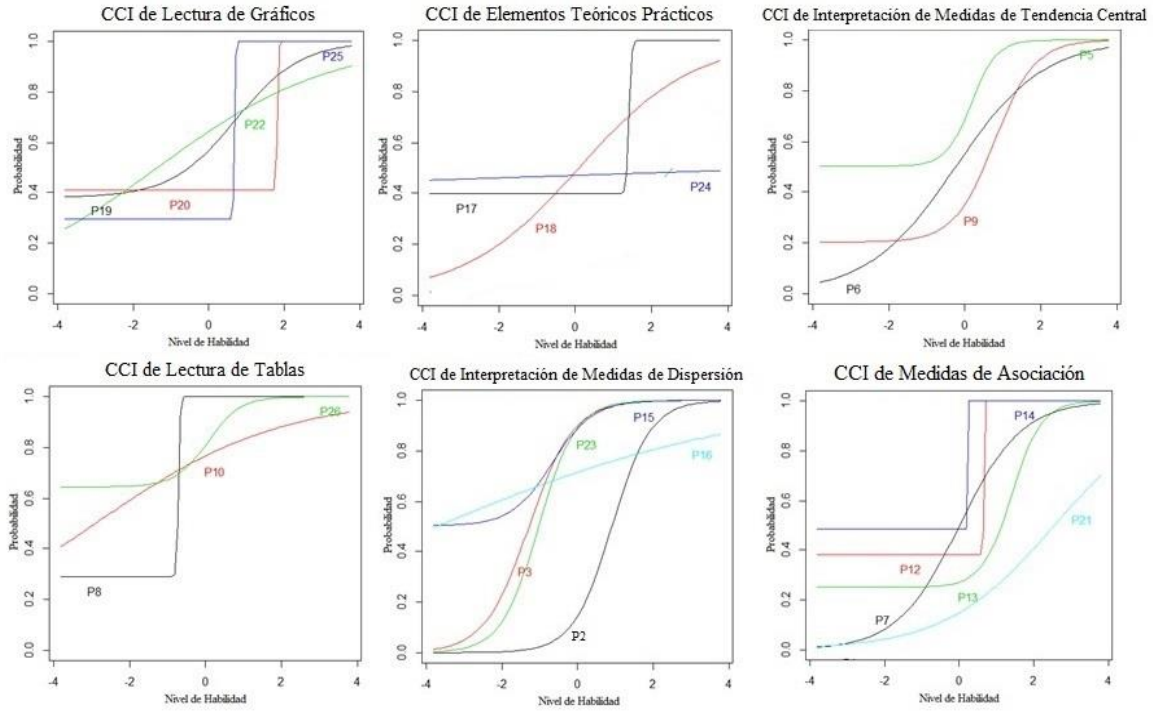


Figura 17. CCI de los ítems en el modelo 3P.

4.4.4.4. Función de Información del Test (FI). En la Figura 18 se observa que la función de información del test posee cuatro picos, los cuales nos indica que el Test Final ajustado en el modelo de tres parámetros proporciona mayor información hacia una región central comprendida entre -1 y 2 pero no es constante en esta tarea, hay 4 puntos donde lo hace muy bien, el máximo se ubica aproximadamente entre 0,5 y 0,8. Como conclusión, esta FI no indica que se este test pueda estimar con precisión niveles de habilidad menores a -1 y mayores a 2 con lo cual, este modelo parece no ser el adecuado para estimar las habilidades de los estudiantes a través del Test Final.

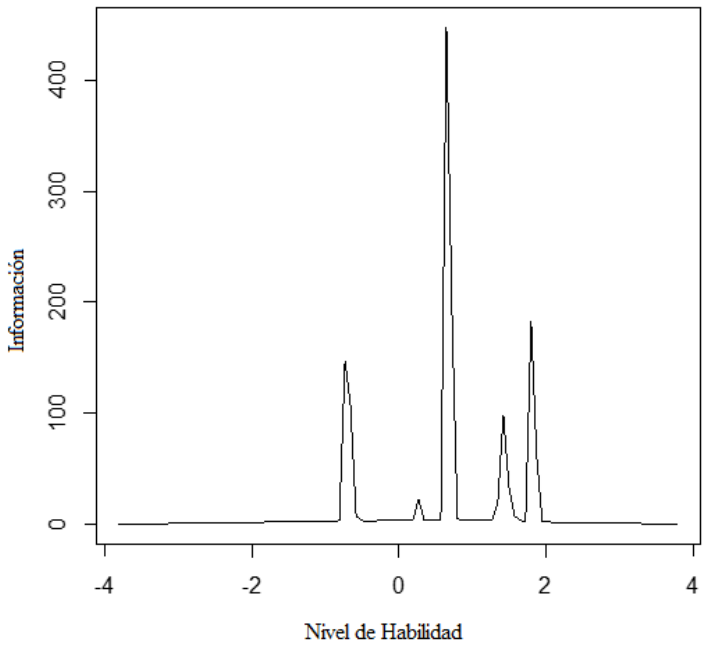


Figura 18. Función de Información del Test en el modelo 3P.

4.4.4.5. Ajuste de la persona. Es importante aquí recordar que este recurso de R tiene por objeto detectar participantes con patrones atípicos de respuesta al estilo de detección de outliers. En general los valores Lz se ubican a ±2 lo cual indicaría que no hay presencia de valores extremos pero en las filas 22 y 40 los valores de Lz son indeterminados. (Ver Tabla 23)

A continuación el autor describe la función de información para cada uno de los tres modelos logísticos, es claro que la información de un test dependerá de los parámetros de discriminación de los ítems (mayores valores en el parámetro a, mayor valor de la información), parámetro de pseudo-azar (menores valores en el parámetro c, mayor información), número de ítems (a mayor longitud, mayor información) y la convergencia entre el nivel de rasgo $\boldsymbol{\theta}$ y los parámetros b de los ítems (cuanto más próximos sean, mayor información).

Tabla 23.

Patrón de respuesta de la persona y su correspondiente habilidad estimada (Lz) – 3P.

	L0	Lz	Pr(<Lz)		L0	Lz	Pr(<Lz)
1	-11,74	0,97	0,83	34	-14,68	-1,49	0,07
2	-16,08	-1,40	0,08	35	-9,53	0,07	0,53
3	-11,27	1,28	0,90	36	-14,42	-0,34	0,37
4	-16,41	-1,55	0,06	37	-10,61	0,58	0,72
5	-12,74	0,37	0,65	38	-11,91	0,56	0,71
6	-11,26	1,24	0,89	39	-10,78	0,88	0,81
7	-11,87	0,73	0,77	40	-10,01	**NaN**	NaN
8	-12,64	0,35	0,64	41	-11,05	0,87	0,81
9	-12,70	0,01	0,50	42	-13,52	-0,40	0,34
10	-12,96	-0,08	0,47	43	-9,45	1,12	0,87
11	-15,98	-1,24	0,11	44	-16,63	-2,00	0,02
12	-13,91	-0,02	0,49	45	-9,80	1,38	0,92
13	-16,11	-1,52	0,06	46	-8,78	1,32	0,91
14	-12,63	0,77	0,78	47	-9,04	-0,57	0,29
15	-10,88	1,20	0,88	48	-17,43	**-2,92**	0,00
16	-16,37	-1,73	0,04	49	-6,94	1,39	0,92
17	-12,90	0,20	0,58	50	-7,45	1,13	0,87
18	-12,31	0,40	0,65	51	-13,27	-1,20	0,12
19	-14,75	-0,58	0,28	52	-11,67	0,03	0,51
20	-13,46	-0,32	0,37	53	-11,73	-0,42	0,34
21	-10,59	1,18	0,88	54	-10,17	-0,26	0,40
22	-12,85	**NaN**	NaN	55	-12,98	-0,25	0,40
23	-12,20	0,37	0,64	56	-8,68	-0,59	0,28
24	-11,81	1,26	0,90	57	-11,12	0,24	0,59
25	-9,18	2,08	0,98	58	-9,89	0,50	0,69
26	-8,90	2,09	0,98	59	-10,81	0,86	0,81
27	-12,04	-0,16	0,44	60	-10,35	-0,53	0,30
28	-13,26	-0,21	0,42	61	-8,56	0,47	0,68
29	-13,17	-0,90	0,18	62	-9,60	0,63	0,74
30	-12,09	0,42	0,66	63	-10,00	0,59	0,72
31	-8,50	1,57	0,94	64	-7,53	0,60	0,72
32	-13,74	-1,94	0,03	65	-2,82	1,35	0,91
33	-7,33	1,20	0,88	66	-3,18	1,12	0,87

4.4.4.6. Ajuste de los ítems. Según los valores p de la Tabla 24, todos los ítems se ajustan bien al modelo 3P excepto tal vez P3 y P6.

Tabla 24.

Ajuste de los ítems al modelo – 3P.

Ítem	X^2	Pr(>X^2)
P2	8,61	0,28
P3	14,40	0,04
P5	6,16	0,52
P6	17,74	0,01
P7	12,85	0,08
P8	0,65	1,00
P9	9,01	0,25
P10	7,27	0,40
P12	11,09	0,13
P13	7,12	0,42
P14	6,98	0,43
P15	4,25	0,75
P16	3,91	0,79
P17	8,83	0,27
P18	5,04	0,65
P19	7,30	0,40
P20	8,75	0,27
P21	4,97	0,66
P22	3,42	0,84
P23	4,92	0,67
P24	10,42	0,17
P25	12,49	0,09
P26	10,01	0,19

4.4.5. Comparación de modelos. La Tabla 25 muestra los estadísticos que permiten analizar la bondad de ajuste de cada uno de los modelos de la TRI a los datos que se disponen por ahora del Test Final. Tras la comparación se concluye que en ausencia de más datos, el modelo 1P sería

el más apropiado para estos datos pero notamos que los valores para el modelo 2P no distan considerablemente.

Tabla 25.

Comparación bondad de ajuste para los modelos 1P, 2P y 3P al Test Final.

Modelo	AIC	BIC	log,Lik
1P	**1815,63**	**1868,18**	-883,81
2P	**1816,88**	**1917,6**	-862,44
3P	1836,77	1987,86	-849,39

4.5. Desempeño en el test

Se calculó el porcentaje de estudiantes (n=66) que eligieron cada opción de respuesta en cada ítem como se muestra en la Tabla 26. Los ítems P2, P13, P17, P20, P21 y P25 tuvieron un porcentaje bajo en la elección de la respuesta correcta en el Test Final, igual comportamiento se observó con los estudiantes en la prueba piloto. También es pertinente identificar que los ítems P20 y P25 miden el dominio de lectura e interpretación de gráficos, los ítems P13 y P21 tratan sobre medidas de asociación, el ítem P2 tiene que ver con interpretación de medidas de dispersión y el ítem P17 se basaba en el uso de elementos teóricos prácticos.

Tabla 26.

Porcentaje de estudiantes (N=66) que eligieron cada opción de selección múltiple en los 27 ítems empleados en el Test Final.

	a	b	c	d	e	N.R
P1	95,45*	1,52	0	3,03		
P2	24,24*	65,15	9,09	0		1,52

P3	10,61	3,03	3,03	81,81*		1,52
P4	0	7,58	87,87*	4,55		
P5	1,52	72,72*	18,18	3,03		4,55
P6	27,27	54,54*	15,15			3,03
P7	22,73	50*	4,55	21,21		1,52
P8	1,52	86,36*	0	10,61		1,52
P9	42,42*	39,39	12,12	4,55		1,52
P10	75,75*	19,70	1,52	3,03		
P11	3,03	12,12	16,67	65,15*		3,03
P12	16,67	10,61	12,12	59,09*		1,52
P13	33,33*	10,61	1,52	53,03		1,52
P14	4,55	0	68,18*	25,76		1,52
P15	84,84*	3,03	3,03	9,09		
P16	1,52	15,15	71,21*	10,61		1,52
P17	1,52	6,06	42,42*	4,55	45,45	
P18	36,36	48,48*	12,12	1,52		1,52
P19	3,03	9,09	59,09*	1,52	25,76	1,52
P20	6,06	0	0	51,52	42,42*	
P21	31,82	9,09	37,88	4,55	16,66*	
P22	3,03	0	19,70	63,63*	10,61	3,03
P23	3,03	78,78*	6,06	10,61		1,52
P24	1,52	25,76	22,73	0	46,96*	3,03
P25	30,30*	48,48	13,63*	7,58		
P26	7,58	81,81*	1,52	9,09		
P27	0	84,84*	4,55	9,09		1,52

Nota. En esta tabla están todos los estudiantes que presentaron el Test Final ya que lo completaron en su totalidad. Los ítems sin resultados presentados para la opción D y E de selección múltiple representan un ítem que no tiene una opción D y E. * indica la respuesta correcta. Adaptada de Ziegler (2014).

Finalmente, en la Figura 19 se analizaron las calificaciones totales de los estudiantes para los 27 ítems del Test Final, el puntaje promedio obtenido fue de 3.32, los puntajes se distribuyen con una variación considerable ($Rango = 3.5; \sigma = 0.71$), solo el 31% de los participantes (20) tuvieron un resultado inferior a 3,0.

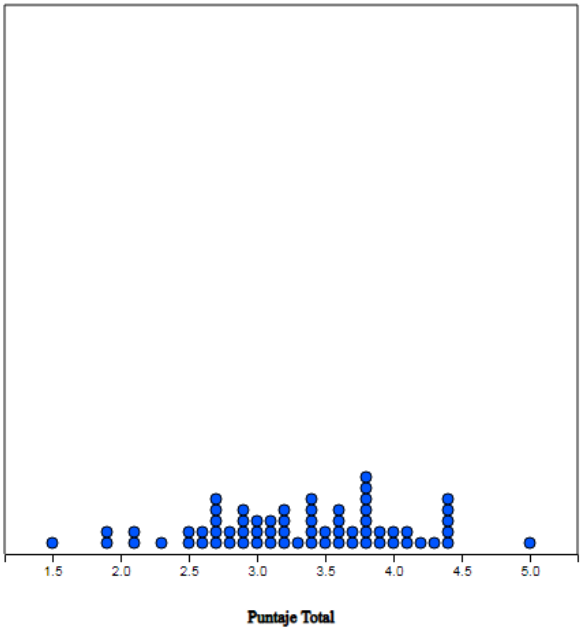


Figura 19. Diagrama de puntos para las puntuaciones totales de los estudiantes en el Test Final.

La Figura 20 muestra el alto grado de asociación lineal entre el nivel de habilidad estimado y la calificación en la escala tradicional (r = 0,99).

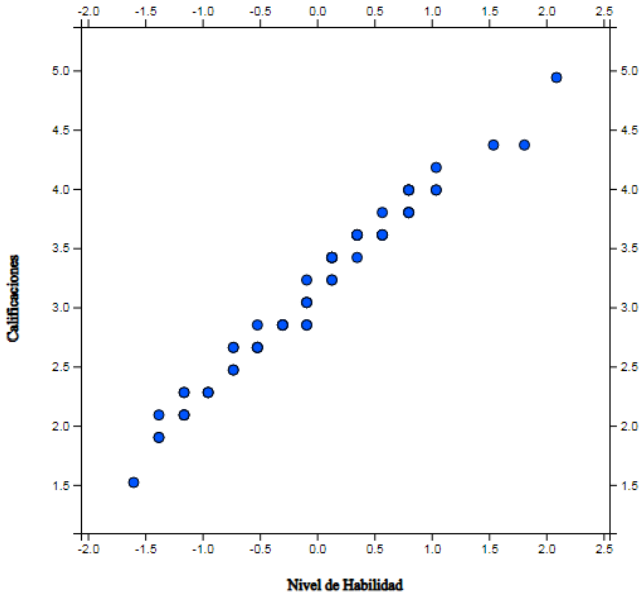


Figura 20. Nivel de Habilidad estimado a partir del modelo de un parámetro vs Calificación (escala de 0 a 5).

4.5.1. Análisis de los individuos con el modelo ajustado. A raíz de que los datos presentaron un mejor ajuste en el modelo 1P se utilizaron los valores de los parámetros de dificultad de los ítems de la Tabla 13 y los niveles de habilidad obtenidos en la Tabla 14.

Se realizó un estudio de los parámetros de dificultad de los ítems por género como se ilustra en la Figura 21 para analizar quienes obtuvieron un mejor razonamiento estadístico si las mujeres o los hombres. De acuerdo con el grafico entre los parámetros de dificultad estimados de los ítems en el género masculino y femenino se presentan diferencias considerables, los casos en que resaltan más estas diferencias están en los ítems P2, P13 y P25 que presentaron una mayor dificultad para las mujeres, con una diferencia de 2.2, 1.57 y 1.91 con respecto al parámetro de dificultad de los hombres. Finalmente se determina que el género masculino tuvo un mejor razonamiento estadístico con respecto al género femenino.

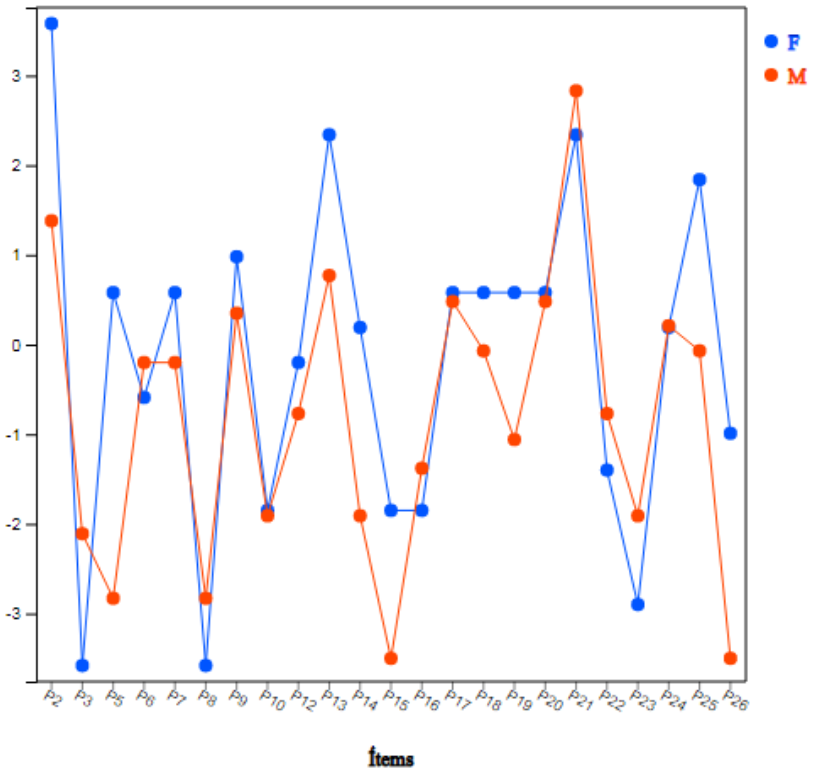


Figura 21. Dificultad de los ítems para el género femenino y masculino.

En la Figura 22 se presenta la gráfica de los parámetros de dificultad de los ítems en los cursos de estadística que nos indica que hubo un mejor razonamiento estadístico en el grupo de ingeniería civil al compararlo con el de matemáticas. Al analizar los resultados se evidencia que el ítem P21 fue el de mayor dificultad para los estudiantes de ambas carreras con diferentes valores en este parámetro.

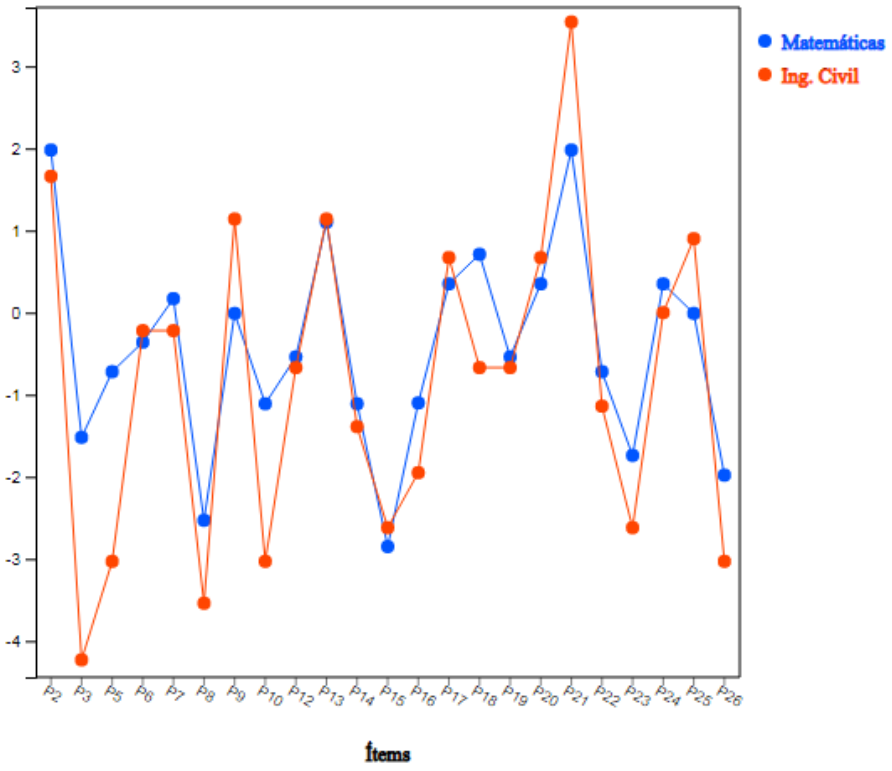


Figura 22. Dificultad de los ítems para cada curso de estadística (Matemáticas e Ingeniería Civil).

Los niveles de habilidad obtenidos en el modelo de ajuste se encuentran en un rango de -1,6 y 2,09. Clasificamos los niveles de habilidad por intervalos como se ilustra en la Tabla 27 y se pudo evidenciar que el 77.27% de los estudiantes (51) tienen un nivel de razonamiento bajo- básico.

Adicionalmente el 90,90% de los individuos (60) que obtuvieron niveles de habilidad entre -2 y 1, presentaron gran dificultad en los ítems del dominio de medidas de asociación, lo cual nos indica que debe haber un refuerzo en este tema en los cursos de estadística que fueron evaluados con el test final.

Tabla 27.

Desempeño de los individuos en el Test Final.

Nivel de Habilidad por Intervalos	N° de estudiantes	Nivel de Razonamiento	Descripción del desempeño
-2 a -1	9	Muy Bajo	Se les dificulta ítems de medidas de asociación, interpretación de medidas de dispersión, interpretación de medidas de tendencia central y elementos teóricos básicos.
-0,9 a 0	23	Bajo	Responden ítems de lectura e interpretación de tablas e interpretación de medidas de dispersión. Se les dificulta ítems de medidas de asociación y lectura e interpretación de gráficos.
0,1 a 1	28	Básico	Responden ítems de interpretación de medidas de tendencia central, lectura e interpretación de tablas e interpretación de medidas de dispersión. Se les dificulta ítems de medidas de asociación.
1,1 a 2	5	Aceptable	Responden ítems de lectura e interpretación de tablas, interpretación de medidas de dispersión, interpretación de medidas de tendencia central y elementos teóricos básicos.
2,1 a 2,5	1	Alto	Responde los ítems de los seis dominios evaluados.

5. Conclusiones

La teoría de respuesta al ítem goza de un interés que crece con el tiempo, desde su aparición en los años sesenta no ha perdido vigencia y por el contrario gracias al desarrollo de los recursos computacionales su estudio y aplicación se ha multiplicado. En este trabajo sólo se estudiaron y aplicaron los modelos clásicos de la TRI pero la literatura más reciente da cuenta de una amplia gama de modelos, entre los que se podrían listar hasta 34 diferentes, así como diferentes opciones computacionales tanto de software libre como comercial. Lo anterior se contrasta con las pocas evidencias de implementación de esta teoría a nivel nacional. Por lo anterior es claro que se debe motivar el estudio de este tema con una proyección hacia los modelos más avanzados, el diseño de Test Adaptativos Computarizados.

En cuanto al banco de ítems se logró recopilar un número importante de ítems de los cuales se examinaron 43 y de ese número se pasó a una versión final del test con 27 ítems, de los cuales se examinaron 23 en la prueba final. Los criterios utilizados para hacer la anterior selección fueron el funcionamiento adecuado del ítem en la prueba piloto entendiendo este como que la respuesta a una pregunta no se concentrara en una única opción de respuesta garantizando así que los distractores cumplen su papel, no se incorporó el análisis de la función de información para cada ítem por limitaciones en cuanto a software. Adicional a estos criterios técnicos, se tuvo en cuenta la longitud de la prueba y la participación de todos los dominios de evaluación con más de una pregunta por cada uno de ellos y con diferentes niveles de dificultad.

Sobre la calibración del test se logró validar de manera aceptable la confiabilidad del test, en cuanto a la unidimensionalidad ésta se pudo validar al reconocer la presencia de testlets en la prueba. La validez fue asumida dado que 23 de los ítems provenían de estudios donde ésta

característica fue estudiada por grupos de expertos, quedaría pendiente por realizar este mismo proceso a los 4 ítems propuestos por las autores de este trabajo. Los parámetros fueron estimados en el paquete ltm de R observándose diferencias importantes en el nivel de dificultad estimado para varios ítems.

A partir de los resultados obtenidos y aplicando los criterios de bondad de ajuste, en el momento se tendrían dos opciones como modelos con buen ajuste a los datos, tanto para 1P como 2P los ítems mostraron buen ajuste a estos modelos, la función de información en ambos casos sugiere que el test es más preciso estimando en los valores centrales del nivel de habilidad. No obstante el AIC para el modelo 1P es levemente inferior al del modelo 2P con lo cual a falta de más datos, el modelo 1P sería el mejor de las tres opciones examinadas. El modelo 3P parece no adecuarse bien a los datos principalmente porque el parámetro de pseudo-azar parece no ser necesario dado que las estimaciones en esta aplicación fueron inferiores a 0.5.

El rendimiento de los estudiantes se puede calificar como aceptable ($\overline{x}$=3.32; s=0.71) sólo una tercera parte de los participantes se ubicó por debajo de la nota mínima aprobatoria que es 3.0; también se destaca que el test logró discriminar a los participantes en todos los niveles de la escala de habilidad (Mínimo= -2.92; Máximo= 2.09). En cuanto a los dominios evaluados se observó un buen desempeño en lectura de tablas e interpretación de medidas de tendencia central, en cambio se perfila como el dominio más débil la interpretación de medidas de asociación y la lectura de gráficas lo cual guarda coherencia con los niveles de dificultad estimados y las CCI presentadas para los modelos 1P y 2P. Destacar también que los ítems que resultaron ser los más difíciles fueron P2 y P21 los cuales hacen parte de los dominios de interpretación de medidas de dispersión y medidas de asociación.

Los niveles de habilidad de los estudiantes estuvieron en un rango de -1,6 y 2,09 presentándose gran acumulación en el intervalo de -0,9 a 1 lo que evidencia que en general las habilidades desarrolladas para analizar datos desde un enfoque descriptivo pueden ubicarse en nivel bajo y básico.

La dificultad de los ítems estuvo comprendida entre -2,95 y 2,58, siendo los ítems de mayor dificultad aquellos que hacían parte del dominio de asociación ya que el 90,90% de los estudiantes (60) que realizaron el test final presentaron falencias en los ítems de este dominio. Los ítems de menor dificultad son aquellos que conforman los dominios de interpretación de medidas de dispersión y lectura e interpretación de tablas pues el 86,36% de los individuos (57) presentaron un buen desempeño en estos ítems.

A pesar de que la recolección de datos abarcó dos semestres académicos no fue posible conseguir un tamaño muestral adecuado, hecho que sin duda afecta los resultados obtenidos por lo cual somos conscientes que no son concluyentes. Uno de los principales inconvenientes para conseguir datos de calidad fue que encontramos que el tema de Estadística Descriptiva es abordado en los inicios de semestre con poca dedicación y siguiendo un enfoque tradicional en el cual se da prioridad a la ejercitación en el uso de las fórmulas dejando de lado la interpretación de datos. Por lo anterior es importante recalcar las condiciones experimentales en que se deben recolectar los datos y la necesidad de controlar factores como los mencionados.

Para dar continuidad a este trabajo se sugiere considerar nuevas aplicaciones del test para ganar mayor precisión en la estimación de los parámetros y poder hacer una mejor evaluación de la bondad de ajuste a los modelos TRI. Otro análisis interesante es contrastar el funcionamiento diferencial de los ítems al discriminar por variables categóricas como el género, programa académico, desempeño en razonamiento cuantitativo, etc.

Finalmente convendría también, avanzar hacia la implementación computacional del test de tal manera que el participante pueda recibir la retroalimentación inmediata, estudiar el efecto de otros factores como el orden de las opciones de respuesta, incluir otros formatos de preguntas que incluso podrían incorporar recursos dinámicos e interactivos.

Referencias bibliográficas

Attorresi, H., Lozzia, G., Abal, F., Galibert, M., y Aguerri, M. (2009). Teoría de Respuesta al ítem: Conceptos Básicos y Aplicaciones para la Medición de Constructos Psicológicos. *Revista Argentina de Clínica Psicológica*. *18*(2), 179-188. Recuperado de http://www.cienciared.com.ar/ra/usr/35/825/racp_xviii_2_pp179_188.pdf

Argibay, J. (2006). Técnicas Psicométricas. Cuestiones de Validez y Confiabilidad. *Subjetividad y Procesos Cognitivos*, (8), 15-33. Recuperado de https://www.redalyc.org/pdf/3396/339630247002.pdf

ARTIST. Recuperado de https://app.gen.umn.edu/artist/

Barajas,A., y Esparza,O. (2010). *Implementación del Modelo Rasch para la estimación de la habilidad algebraica de los estudiantes de primer semestre de ciencias e ingeniería de la Universidad Industrial de Santander* (Tesis Pregrado). Universidad Industrial de Santander, Bucaramanga, Colombia. Recuperado de http://tangara.uis.edu.co/biblioweb/tesis/2010/134776.pdf

Carmines, E., y Zeller, R. (1979). *Reiability and validity assessment*. Londres: Sage. Recuperado de https://www.researchgate.net/publication/260763618

CAOS. Recuperado de https://apps3.cehd.umn.edu/artist/articles/AERA_2006_CAOS.pdf

Carvajal, D., Méndez, H., y Torres, M. (2016). *Análisis de la confiabilidad y de algunos parámetros psicométricos de un test realizado en el Colegio Vista Bella de la ciudad de Bogotá* (Tesis Posgrado). Fundación Universitaria los Libertadores, Bogotá D.C., Colombia. Recuperado de

https://repository.libertadores.edu.co/bitstream/handle/11371/620/Carvajal%C3%81lzateDiegoEliezer.pdf?sequence=2&isAllowed=y

Contreras, J., Cañadas, G., Gea, M., y Arteaga, A. (2013). *Actas de las Jornadas Virtuales en Didáctica de la Estadística, Probabilidad y Combinatoria*. Universidad de Granada, pp. 391-396. Recuperado de http://estadis.net/3/documentos/ACTAS/2%20Comunicacion%2040.pdf

DelMas, R., Garfield, J., y Ooms, A. (2005). *Using Assessment Items to Study Students' Difficulty Reading and Interpreting Graphical Representations of Distributions.* Recuperado de https://www.causeweb.org/cause/archive/artist/articles/SRTL4_ARTIST.pdf

DelMas, R., Garfield, J., Ooms, A., y Chance, B. (2006). Assessing Students' Conceptual Understanding after a First Course in Statistics. *Annual Meetings of The American Educational Research Association*, San Francisco, EU. Recuperado de https://apps3.cehd.umn.edu/artist/articles/AERA_2006_CAOS.pdf

DelMas, R., Garfield, J., Ooms, A., y Chance, B. (2007). Assessing Students' Conceptual Understanding after a First Course in Statistics. *Statistics Education Research Journal,* 6(2), 28-58. Recuperado de https://www.stat.auckland.ac.nz/~iase/serj/SERJ6(2)_delMas.pdf

EMAR. (2018). ¿Qué es la teoría de respuesta al ítem?. *Racionalidad Ltda*. Recuperado de https://racionalidadltda.wordpress.com/2018/04/13/que-es-la-teoria-de-respuesta-al-item/

Ferrando, J. (1996). Evaluación de la Unidimensionalidad de los ítems mediante análisis factorial. Recuperado de http://www.psicothema.com/pdf/38.pdf

Garfield, J., DelMas, B., Chance, B., y Ooms, A. (2006). *ARTIST Scale: Data Collection.* Recuperado de https://app.gen.umn.edu/artist/

Garfield, J., DelMas, B., Chance, B., y Ooms, A. (2006). *ARTIST Scale: Bivariate Data, Quantitative.* Recuperado de https://app.gen.umn.edu/artist/

Garfield, J., DelMas, B., Chance, B., y Ooms, A. (2006). *ARTIST Scale: Data Representation.* Recuperado de https://app.gen.umn.edu/artist/

Garfield, J., DelMas, B., Chance, B., y Ooms, A. (2006). *ARTIST Scale: Measures of Center.* Recuperado de https://app.gen.umn.edu/artist/

Garfield, J., DelMas, B., Chance, B., y Ooms, A. (2006). *ARTIST Scale: Measures of Spread.* Recuperado de https://app.gen.umn.edu/artist/

Goforth, Ch. (2015). Using and Interpreting Cronbach's Alpha. *University of Virginia Library*. Recuperado de https://data.library.virginia.edu/using-and-interpreting-cronbachs-alpha/

GRE. *Official GRE Quantitative Reasoning Practice Questions*. (2014). Estados Unidos y otros países: ETS. Recuperado de http://46.100.53.74/IdeaWeb/Files/Data/Library/20165231053569.pdf

Hidalgo, M., French, B. (2016). Una Introducción Didáctica a la Teoría de Respuesta al ítem para Comprender la Construcción de Escalas. *Revista de Psicología Clínica con Niños y Adolescentes*. *3*(2), 13-21. Recuperado de https://dialnet.unirioja.es/servlet/articulo?codigo=5590670

ICFES. *Banco de Preguntas de Matemáticas*. Recuperado de http://www.mentesenblanco-razonamientoabstracto.com/icfes-banco-de-preguntas/matematicas.pdf

Leenen, I. (2014). Virtudes y limitaciones de la teoría de respuesta al ítem para la evaluación educativa en las ciencias médicas. *Investigación en Educación Médica*. 3(9), 40-55. Recuperado de http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S20075057201400010007

Liu, Y. y Maydeu-Olivares, A. (2012). Local dependence diagnostics in IRT modeling of binary data. Recuperado de http://www.ub.edu/gdne/documents/local_dependence_epm12.pdf

Morgado, C., y Neusa, M. (2011). *Análisis de las Olimpiadas Regionales de Matemáticas UIS implementando el modelo Rasch para los años 2009 y 20010* (Tesis Pregrado). Universidad Industrial de Santander, Bucaramanga, Colombia. Recuperado de http://tangara.uis.edu.co/biblioweb/tesis/2011/137801.pdf

Muñiz, J. (1997). Introducción a la teoría de respuesta a los ítems. Madrid: Ediciones Pirámide. Recuperado de http://innoevalua.us.es/files/irt.pdf

Muñiz, J. (1998). La Medición de lo Psicológico. *Psicothema.* 10(1), 1-21. Recuperado de http://www.psicothema.com/psicothema.asp?id=138

Nava, M. (1994). Teoría Clásica de los Tests versus Teoría de Respuesta al Ítem. *Psicológíca*, 15, 175-208. Recuperado de http://www2.uned.es/490015/CV/TCTTRI94.pdf

Olea, J., y Ponsoda, V. (2002). *Test Adaptativos Informatizados*. Universidad Autónoma de Madrid. Recuperado de https://www.researchgate.net/publication/265040034_TEST_ADAPTATIVOS_INFORMATIZADOS

Paek, I., y Cole, K. (2020). *Using R for Item Response Theory Model Applications.* Nueva York, EU. Abingdon, Reino Unido: Routledge.

Pérez, J. (2011). Módulos de Medición: *Desarrollos actuales, supuestos, ventajas e inconvenientes*, 1, 1-31. Universidad de Sevilla. Recuperado de http://innoevalua.us.es/files/irt.pdf

Reckase, J. (2018). Cursillo: *An Introduction to Computarized Adaptive Testing*. Memorias del XXVIII Simposio Internacional de Estadística, Bucaramanga, Colombia.

Rojas, M., Manríquez, G., Gatica, Y., y Salcedo, P. (2004). Curso de UML Multiplataforma Adaptativo Basado en la Teoría de Respuesta al ítem. Universidad de Concepción. *Revista Ingeniería Informática, 10*. Recuperado de http://inf.udec.cl/~revista/ediciones/ edicion10/psalcedo01.pdf

Sabbag, A., Garfield, J., y Zieffler, A. (2018). *Assessing Statistical Literacy and Statistical Reasoning: The Reali Instrument.* Recuperado de https://iaseweb.org/documents/ SERJ/SERJ17(2)_Sabbag.pdf

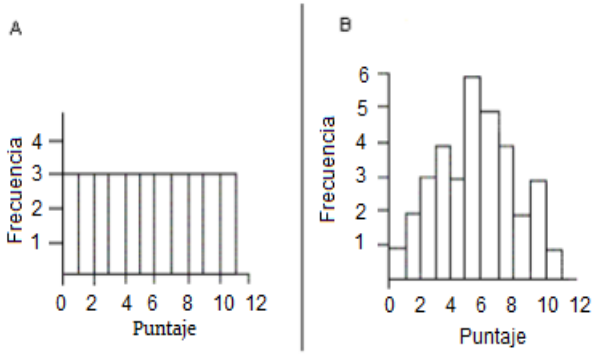
Streefkerk, R. (2019). Internal vs External validity. *Scribbr*. Recuperado de https://www.scribbr.com/methodology/internal-vs-external validity/#:~:text=Internal%20validity%20refers%20to%20the,other%20situations%2C%20groups%20or%20events.

Tshering, G. (2006). *IRT in Item Banking, Study of DIF Items and Test Construction: Item Response Theory in Item Banking, Study of Differentially Functioning Items and Test Construction* (Tesis Posgrado). University of Twente, Países Bajos. Recuperado de https://www.researchgate.net/publication/279492212

Ziegler, L. (2014). *Reconceptualizing* Statistical *Literacy: Developing an Assessment for the Modern Introductory Statistics Course* (Tesis doctoral). Minnesota University, Minnesota, EU.

Apéndices

Apéndice A. Banco de Ítems.

ÍTEM	ENUNCIADO	TEST
P1	Para un trabajo de Estadística un estudiante va a recolectar datos entre los estudiantes que llegan a la Universidad en vehículo particular. Una de las variables que le interesa estudiar es la marca del vehículo (Renault, Mazda, Hiundai, Honda, Mazda, Ford, Chevrolet). ¿Qué tipo de variable es la que va a medir este estudiante? a. Categórica b. Cuantitativa c. Continua d. Discreta	ARTIST
P2	Para este par de gráficas, determine cuál gráfica tiene la más alta desviación estándar (No se requiere el cálculo exacto para responder).  a. A tiene mayor desviación estándar que B. b. B tiene mayor desviación estándar que A. c. Ambos gráficos exhiben la misma desviación estándar. d. Como B tiene más individuos, los puntajes varían más.	ARTIST

P3	Treinta estudiantes del curso de Estadística tomaron un quiz cuyo máximo puntaje era 30 puntos. La desviación estándar observada fue de 1 punto. ¿Cuál de los siguientes enunciados es la interpretación más adecuada de la desviación estándar? a. Todos los puntajes tienen un punto de diferencia b. La diferencia entre el puntaje más alto y el más bajo es de 1 punto c. La diferencia entre el cuartil 3 y el cuartil 1 es de 1 punto d. La distancia típica de un puntaje a la media es de 1	ARTIST
P4	Una clase de 30 estudiantes tomo un quiz de 15 preguntas, con cada pregunta ganaba un punto. La desviación estándar para la distribución resultante es cero. Bajo esta información Uds. conoce que: a. Más de la mitad de los puntajes estuvieron por encima de la media. b. Imposible observar ese valor; se debió cometer un error aritmético. c. Cada persona correctamente respondió el mismo número de preguntas. d. La media, la mediana y la moda deben ser todas cero.	ARTIST
P5	Se aplica un test a 100 estudiantes y se determina la mediana de los puntajes. Luego de calificar los test, usted nota que los diez estudiantes con las notas más altas lo hicieron excepcionalmente bien, y decide dar un premio a estos 10 estudiantes de 5 puntos adicionales. La mediana de los nuevos puntajes será _____ que la distribución de los puntajes originales. a. Más baja que b. Igual a c. Más alta que d. Dependiendo de la asimetría será tan alta como o tan baja como	ARTIST

<table>
<tr>
<td>P6</td>
<td>La distribución del ingreso entre los individuos del 1% superior en USA es fuertemente sesgada a la derecha. En 2017, las dos medidas de tendencia central para los individuos del 1% superior fueron $330.000 y 675.000. ¿Cuál número representa el ingreso medio y cuál número representa el ingreso mediano en el 1% superior en cuanto a ingresos?

a. $330.000 es la media y $675.000 es la mediana.
b. $330.000 es la mediana y $675.000 es la media.
c. No hay suficiente información para responder la pregunta.
A continuación, se presentan los puntajes obtenidos por dos grupos en una prueba de aptitud. ¿Qué concluiría sobre el rendimiento de estos dos grupos?</td>
<td>ARTIST</td>
</tr>
<tr>
<td>P7</td>
<td>A continuación, se presentan los puntajes obtenidos por dos grupos en una prueba de aptitud. ¿Qué concluiría sobre el rendimiento de estos dos grupos?

<table>
<tr><th>A</th><th>B</th></tr>
<tr><td>65</td><td>70</td></tr>
<tr><td>73</td><td>62</td></tr>
<tr><td>52</td><td>50</td></tr>
<tr><td>48</td><td>72</td></tr>
<tr><td>74</td><td>66</td></tr>
<tr><td>61</td><td>66</td></tr>
<tr><td>60</td><td>51</td></tr>
<tr><td>75</td><td>48</td></tr>
<tr><td>47</td><td>46</td></tr>
<tr><td>73</td><td>50</td></tr>
<tr><td>48</td><td>44</td></tr>
<tr><td>53</td><td>73</td></tr>
<tr><td>46</td><td>61</td></tr>
<tr><td>60</td><td>66</td></tr>
<tr><td>56</td><td>46</td></tr>
<tr><td>60</td><td>71</td></tr>
<tr><td>50</td><td>66</td></tr>
<tr><td>60</td><td>55</td></tr>
</table>

<table>
<tr><th></th><th>A</th><th>B</th></tr>
<tr><td>n</td><td></td><td></td></tr>
<tr><td>Media</td><td></td><td></td></tr>
<tr><td>Mediana</td><td>60</td><td>61,5</td></tr>
<tr><td>Moda</td><td>60</td><td>66</td></tr>
<tr><td>Desviación</td><td></td><td></td></tr>
<tr><td>Varianza</td><td>96,879</td><td>103,82</td></tr>
<tr><td>Asimetría</td><td>0,386</td><td></td></tr>
<tr><td>Curtosis</td><td></td><td>-1,65</td></tr>
<tr><td>Valor mínimo</td><td>46</td><td>44</td></tr>
<tr><td>Valor máximo</td><td>75</td><td>73</td></tr>
</table>
</td>
<td>Examen de Estadística Descriptiva</td>
</tr>
</table>

<table>
<tr>
<td>P8</td>
<td>Una empresa ha hecho un estudio para determinar qué tan conocido es el producto que ofrece. Para este estudio realizaron encuestas dividiendo la población encuestada en tres grupos. Los resultados fueron los siguientes:

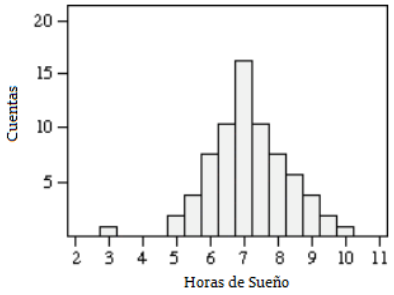
<table>
<tr><th>Grupo</th><th>Total de personas encuestadas</th><th>Cantidad de personas que conocen que existe el producto pero no lo usan</th><th>Cantidad de personas que conocen y usan el producto</th></tr>
<tr><td>I</td><td>200</td><td>110</td><td>70</td></tr>
<tr><td>II</td><td>500</td><td>250</td><td>220</td></tr>
<tr><td>III</td><td>150</td><td>120</td><td>20</td></tr>
</table>

Según las expectativas de la empresa, se fijó que el producto permanecería en el mercado si el 60% de la población hace uso de él. A partir de los resultados del estudio es más probable que

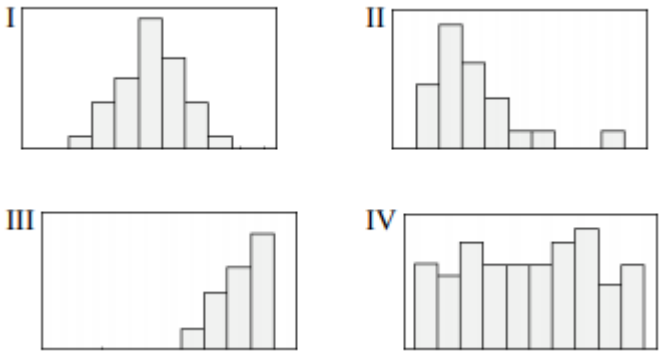
a. El producto continúe en el mercado, porque en todos los grupos la cantidad de personas que no usan el producto es menor que la cantidad de los que lo usan
b. El producto no continúe en el mercado, porque sólo 31 de cada 85 personas encuestadas usan el producto
c. El producto continúe en el mercado, porque sólo 6 de cada 85 personas encuestadas no conocen el producto
d. El producto no continúe en el mercado, porque el porcentaje de encuestados en el grupo III que usa el producto es aproximadamente el 2,3% de los encuestados</td>
<td>ICFES</td>
</tr>
<tr>
<td>P9</td>
<td>En Norteamérica, una encuesta a nivel nacional aplicada a propietarios de perros grandes indicó que el costo promedio en el primer año era de 1.700 dólares. Cuál de las siguientes es la mejor interpretación de la media:

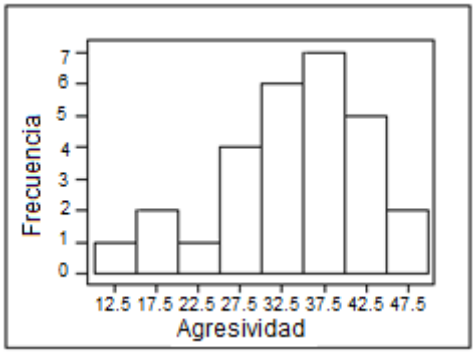
a. Para los propietarios de perros grandes de esta muestra, el costo promedio durante el primer año es de 1.700 dólares.
b. Para los propietarios de perros grandes de esa población, el costo promedio durante el primer año es de 1.700 dólares.
c. Aproximadamente la mitad de los valores dados por los propietarios de perros grandes en esta muestra se ubican por encima de 1.700 y la otra mitad estuvo por debajo de 1.700.
d. Para muchos propietarios de perros grandes, el costo durante el primer año es de 1.700 dólares.</td>
<td>BLIS-2</td>
</tr>
</table>

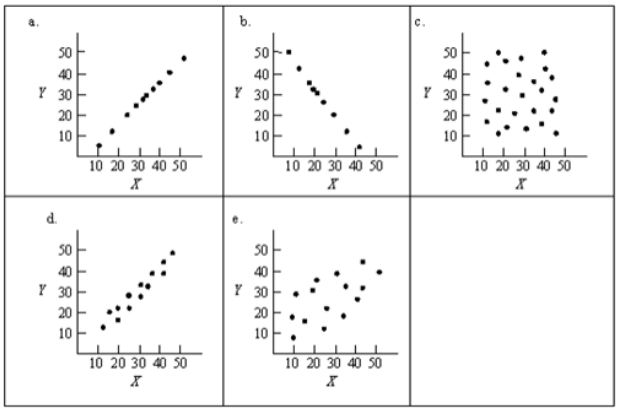
<table>
<tr>
<td>P10</td>
<td>Para tomar la decisión de construir una plaza de mercado en el barrio Los Rosales, la Junta de Acción Comunal desea contar con el apoyo de la mayoría de las familias que allí viven. Para determinar qué quiere la mayoría, realizaron un sondeo en el que preguntaron: "¿Cree usted que sería de beneficio para el sector la construcción de una plaza de mercado?". Los resultados se muestran en la siguiente tabla:
<table>
<tr><th>Respuesta</th><th>N° de Familias</th></tr>
<tr><td>Si</td><td>225</td></tr>
<tr><td>No</td><td>150</td></tr>
<tr><td>Esta inseguro</td><td>75</td></tr>
<tr><td>No respondió</td><td>300</td></tr>
</table>
La Junta de Acción Comunal se inclinó por NO construir una plaza de mercado, debido a que los resultados del sondeo muestran que:

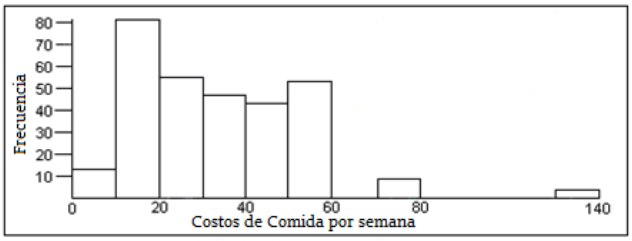
a. El 70% de familias encuestadas no respondió afirmativamente.
b. La mitad de familias encuestadas estuvieron inseguras o no respondieron la encuesta.
c. El número de familias que respondieron "sí", supera a quienes respondieron negativamente en un 50%.
d. El número de familias que respondieron "no" es el doble de las que están inseguras.</td>
<td>ICFES</td>
</tr>
<tr>
<td>P11</td>
<td>La siguiente gráfica muestra una distribución de las horas de sueño del día anterior entre un grupo de estudiantes universitarios. Seleccione la afirmación que ofrece la descripción más completa de la gráfica en una forma que demuestre un entendimiento de cómo en Estadística se describen e interpreta la distribución de una variable.


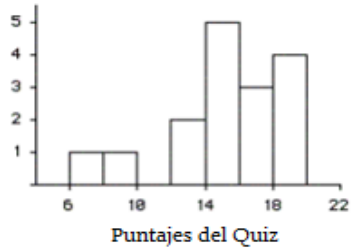
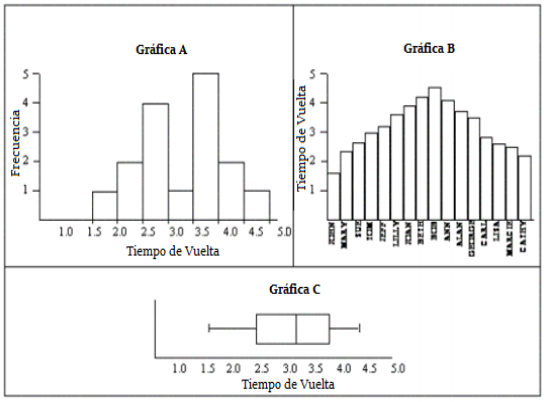
a. Las barras van de 3 a 10, incrementando la altura en 7 y decreciendo en 10. La barra más alta se ubica en 7. Hay un vacío entre tres y cinco.</td>
<td>CAOS</td>
</tr>
</table>

	b. La distribución es normal (simétrica) con una media de 7 y una desviación estándar de 1. c. Muchos estudiantes consideran que tendrán suficiente sueño en la noche, pero algunos estudiantes duermen más y algunos duermen menos. Sin embargo, un estudiante debe haber estado despierto gran parte de la noche por eso durmió tan pocas horas. d. La distribución de las horas de sueño es simétrica, bien curveada con valor atípico en 3.	
P12	**Los ítems P12, P13, P14 se responden a partir de los siguientes datos.** Escoja la gráfica que mejor representa la variable indicada en cada pregunta, no hay restricciones así que podría repetir respuesta.  Los últimos dígitos de los números de teléfono tomados de un directorio telefónico. a. I b. II c. III d. IV	CAOS
P13	El conjunto de los pesos promedio (medido en libras) medidos mensualmente durante dos años para un grupo de adultos saludables. Encaja esta descripción con un histograma de los presentados. a. I b. II c. III d. IV	CAOS

P14	Un conjunto de puntajes en un quiz donde el quiz fue muy fácil de responder. a. I b. II c. III d. IV	CAOS
P15	Para la lista de datos seleccione la mejor estimación para la desviación estándar. La media es 50. No se requieren los cálculos para poder responder. Lista A: 49, 51, 49, 51, 49, 51, 49, 51, 49, 51 a.1 b. 2 c. 5 d. Ninguna	ARTIST
P16	Para la lista de datos seleccione la mejor estimación para la desviación estándar. La media es 50. No se requieren los cálculos para poder responder. Lista B: 31, 36, 48, 50, 50, 53, 54, 56, 60, 62 a.1 b. 3 c. 8 d. 20	ARTIST
P17	Una prueba que mide la tendencia agresiva fue aplicada a un grupo de chicos adolescentes quienes fueron miembros de una pandilla. La prueba califica con puntajes de 10 a 60, aquí un alto puntaje indica más agresividad. El histograma representa el resultado para 28 chicos. ¿Cuáles dos medidas podrían ser más apropiadas para describir el centro y la dispersión de esta distribución?  a. Rango y media b. Media y mediana c. Mediana y Rango intercuartílico d. Moda y rango e. Media y desviación estándar	ARTIST

P18	Un estudiante estuvo estudiando la relación entre cuánto dinero gastan los estudiantes en comida y en entretenimiento por semana. Basado en una muestra de 270 estudiantes, él calculó un coeficiente de correlación para estas dos variables y obtuvo r=0.013. ¿Cuál de las siguientes sería la mejor interpretación? a. Este bajo coeficiente de correlación de 0.013 indica que no hay relación entre estas dos variables. b. No hay relación lineal, pero puede haber una relación no lineal. c. Hay cierto nivel de asociación lineal. d. Las variables no son proporcionales.	ARTIST
P19	**Los ítems P19 – P20 se responden a partir de los siguientes datos.**  Seleccione el gráfico de dispersión que muestra una correlación baja. a. a b. b c. c d. d e. e	ARTIST
P20	Seleccione el gráfico de dispersión que muestra una correlación de 0.6 a. a b. b c. c d. d e. e	ARTIST

P21	A continuación, se presentan las calificaciones del primer parcial de Estadística: 17, 13, 8, 15, 14, 11, 13, 18, 19, 20, 14, 17, 16, 18, 12, 16, 14, 15, 19, 17, 2, 4, 19, 10, 6, 18. Si un estudiante obtuvo una calificación de 14 puntos, ¿Cómo fue su rendimiento en relación con el grupo?	Examen de Estadística Descriptiva
P22	A continuación, se muestra el resultado de una encuesta a una muestra aleatoria tomada entre estudiantes de cierta Universidad. A continuación, se muestra la distribución del presupuesto semanal entre estudiantes. Un estudiante afirma que la distribución del costo en comida básicamente corresponde a una curva simétrica con un valor extremo (outlier). ¿Cómo respondería Uds.?  a. De acuerdo, esta distribución luce muy simétrica si ignoramos el valor extremo. b. De acuerdo, la mayoría de distribuciones son simétricas (normales). c. En descuerdo, ésta gráfica luce más sesgada a la izquierda. d. En descuerdo, ésta gráfica luce más sesgada a la derecha. e. En desacuerdo, ésta gráfica luce más bimodal.	ARTIST
P23	Considere dos poblaciones de una misma región. Ambas poblaciones tienen el mismo tamaño (22.000). La Población 1 está conformada mayoritariamente por estudiantes de una universidad mientras que la Población 2 se integra principalmente por los residentes de una pequeña ciudad. Considere la variable Edad. ¿Cuál población podría más probablemente tener la mayor desviación estándar? a. Población 1 b. Población 2 c. Podrían tener la misma desviación estándar d. No hay suficiente información.	ARTIST

P24	La nota en un quiz fue calculada como el número de respuestas correctas. A continuación, se muestra la gráfica de las notas obtenidas en este quiz. ¿Cuántas de las notas están sobre 15?  a. 6 b. 7 c.12 d.13 e. No puede ser determinado	ARTIST
P25	Un club de atletismo tiene su propia pista y registra con precisión el record de cada miembro en su mejor vuelta, así pueden hacer comparaciones con sus compañeros. A continuación, se muestran las gráficas de estos datos, ¿Cuál de ellas le permite a Ud. fácilmente ver la forma de la distribución de los tiempos de carrera mencionados?  a. A b. B c. C d. Todas las anteriores.	ARTIST

Los ítems P26 – P27 se responden a partir de los siguientes datos.

Los datos corresponden a una encuesta que indagó la opinión por el servicio de residencias universitarias.

	Positiva	**Negativa**	**Neutral**	Total
Hombres	424	303	131	858
Mujeres	295	675	121	1091
Total	719	978	252	1949

% Total	0,218	0,156	0,067
	0,151	0,346	0,062
% Fila	0,494	0,353	0,153
	0,27	0,619	0,111
% Columna	0,59	0,31	0,52
	0,41	0,69	0,48

P26	Con relación al valor 675 de la anterior tabla, una interpretación válida es: a. Entre las mujeres, el 33% tiene una opinión negativa b. Cerca del 33% de los encuestados son mujeres con opinión negativa c. De la opinión negativa, el 33% corresponde a mujeres d. El porcentaje de opinión negativa es mayor entre las mujeres debido a que hay más mujeres en la muestra	Creación Propia
P27	¿Cuál de las siguientes es una conclusión válida sobre la asociación entre el género y la opinión sobre el servicio de residencias (abajo aparecen algunos porcentajes calculados) a. Hay asociación entre las variables género y opinión por ello se observa que es más probable que los hombres tengan una opinión negativa del servicio de residencias. b. Hay asociación entre las variables género y opinión, por ello se observa que es más probable que las	Creación Propia

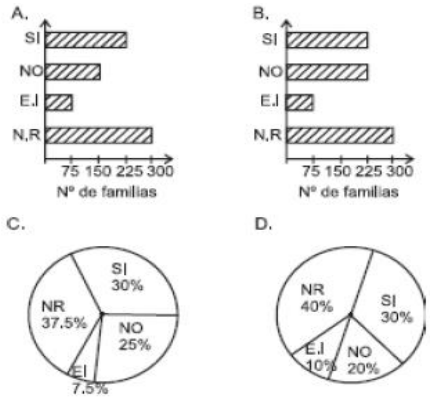
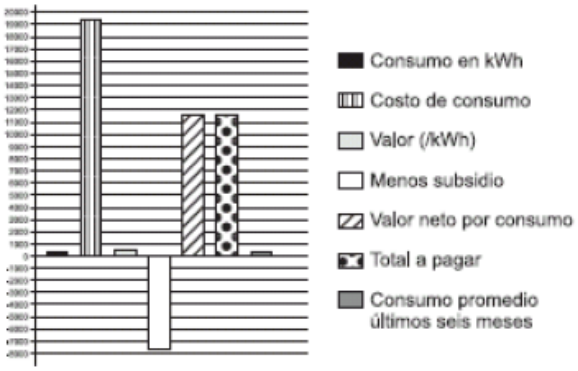
<table>
<tr><td></td><td>mujeres tengan una opinión negativa del servicio de residencias.

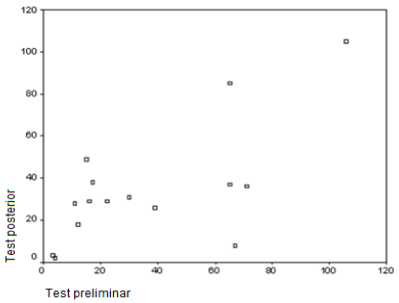
c. No hay asociación entre género y opinión.

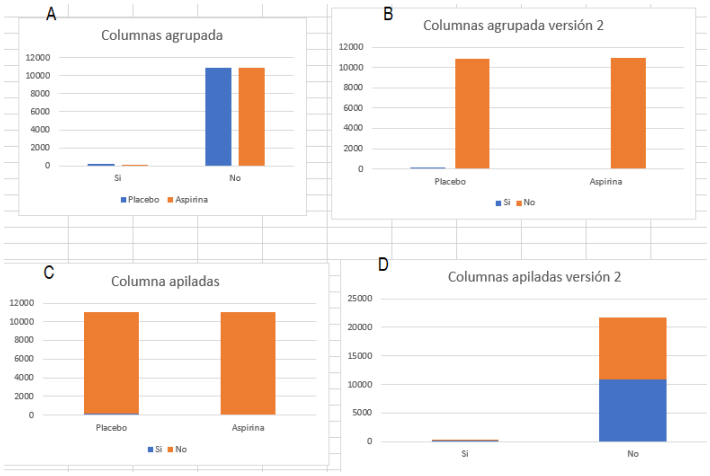
d. Sólo se observa el efecto del género en la opinión negativa.</td><td></td></tr>
<tr><td>P28</td><td>Para un trabajo de Estadística un estudiante va a recolectar datos entre los estudiantes que llegan a la Universidad en vehículo particular. Una de las variables que le interesa estudiar es la marca del vehículo (Renault, Mazda, Hiundai, Honda, Mazda, Ford, Chevrolet). Otro objetivo del trabajo es ver si la marca del vehículo puede predecir el número de comparendos que el conductor recibe en un año. ¿Identifique la variable dependiente en este estudio?

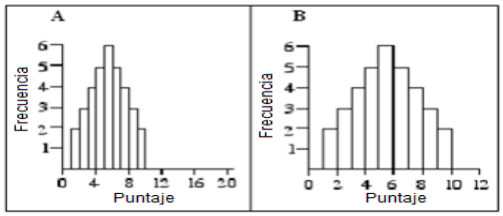
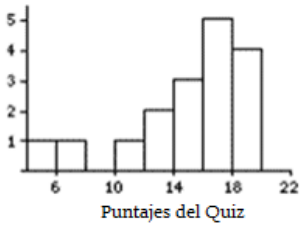
a. Número de estudiantes con vehículo
b. Marca del vehículo
c. Número de comparendos
d. Número promedio de comparendos en el último año</td><td>ARTIST</td></tr>
<tr><td>P29</td><td>Para tomar la decisión de construir una plaza de mercado en el barrio Los Rosales, la Junta de Acción Comunal desea contar con el apoyo de la mayoría de las familias que allí viven. Para determinar qué quiere la mayoría, realizaron un sondeo en el que preguntaron: "¿Cree usted que sería de beneficio para el sector la construcción de una plaza de mercado?". Los resultados se muestran en la siguiente tabla:

<table><tr><td>Respuesta</td><td>N° de Familias</td></tr><tr><td>Si</td><td>225</td></tr><tr><td>No</td><td>150</td></tr><tr><td>Esta inseguro</td><td>75</td></tr><tr><td>No respondió</td><td>300</td></tr></table>
Un gráfico que se podría presentar a los habitantes del barrio, sobre los resultados del sondeo, es</td><td>ICFES</td></tr>
</table>

		
P30	A la casa que comparten cinco jóvenes ha llegado la factura de cobro del servicio de energía correspondiente al consumo del mes de septiembre. Entre la información que aparece en la factura se encuentra la siguiente: Consumo promedio últimos Seis meses en kWh Consumo en (kWh) 110 Valor (/kWh) 175,0952 Costo de consumo 19 260 Menos subsidio -7 704 Valor neto por consumo 11 556 Ajuste decena 4 Total, a pagar 11 560 Uno de los jóvenes ha decidido mostrar a sus compañeros la siguiente representación gráfica de la información proporcionada en la factura 	ICFES

	Uno de los jóvenes, al analizar la gráfica, hace la observación de que no debe presentarse así, puesto que a. En la gráfica se relaciona correctamente la información de la factura, sin embargo, para facilitar la lectura sería más conveniente organizar las barras por tamaño b. La gráfica está mal construida porque la barra que indica subsidio no debería corresponder a un valor negativo ya que es un ahorro y no un gasto c. No es posible relacionar todos los datos de la factura en una gráfica como ésta, porque la escala numérica no puede asociarse a pesos y kWh simultáneamente d. No es posible que la gráfica sea correcta porque el total a pagar no puede ser menor que el costo del consumo.	
P31	En cierta asignatura, se aplica primero un test de conocimientos previos (pretest) y luego de estudiado el tema un test de conocimiento posterior (postest).  Se desea estudiar si hay relación entre estas dos pruebas. Se sabe que el puntaje en el pretest de Jhon es 5 y su puntaje en el postest es 100. Si esta corrección es hecha a los datos y un nuevo coeficiente de correlación es calculado, ¿Cómo será la nueva correlación comparada respecto del coeficiente de correlación calculado inicialmente? a. El valor absoluto de la nueva correlación podría ser más pequeña que el valor absoluto de la correlación original.	ARTIST

<table>
<tr><td></td><td>b. El valor absoluto de la nueva correlación podría ser más grande que el valor absoluto de la correlación original.
c. El valor absoluto de la nueva correlación podría ser el mismo que el valor absoluto de la correlación original.
d. Es imposible predecir cómo será este nuevo coeficiente de correlación.</td><td></td></tr>
<tr><td>P32</td><td>La siguiente tabla cruza las variables Medicamento preventivo vs Infarto.
<table><tr><td></td><td>Si</td><td>No</td></tr><tr><td>Placebo</td><td>189</td><td>10845</td></tr><tr><td>Aspirina</td><td>104</td><td>10933</td></tr></table>
¿Cuál de las siguientes representaciones utilizaría usted para presentar la información más relevante sobre este tema?
 
</td><td>Creación Propia</td></tr>
</table>

P33	Para el siguiente par de gráficas, determine cuál gráfica tiene la más alta desviación estándar (No se requiere el cálculo exacto para responder).  a. A tiene mayor desviación estándar que B. b. B tiene mayor desviación estándar que A. c. Ambos gráficos tienen la misma desviación estándar. d. Como tienen la misma cantidad de individuos, luce más homogénea la distribución A.	ARTIST
P34	Un profesor de Matemáticas aplicó una prueba de 15 preguntas. Al calificar usó el siguiente sistema: cada pregunta respondida correctamente vale un punto, preguntas sin respuesta valen 0 puntos y si la respuesta a una pregunta es incorrecta resta un punto. Una vez calificadas las pruebas, el profesor calcula la desviación estándar y obtiene un valor de -2.30. ¿Qué podemos decir al respecto de este valor? a. Hay un error en el cálculo de la desviación estándar b. Muchos estudiantes obtuvieron calificaciones bajas c. Muchos estudiantes obtuvieron calificaciones por debajo de la media d. Ninguna de las anteriores	ARTIST
P35	La gráfica a continuación muestra los puntajes en un quiz, ¿Cuál estimador de la media y la mediana sería el más posible? 	ARTIST

<table>
<tr>
<td></td>
<td>a. Mediana = 13.0 y Media = 12
b. Mediana = 14.0 y Media = 15
c. Mediana = 16.0 y Media = 14.3
d. Mediana = 16.5 y Media = 16.2</td>
<td></td>
</tr>
<tr>
<td>P36</td>
<td>Algunos estudiantes de una universidad recogieron información acerca del número de hombres y mujeres que nacieron en un hospital durante 2 semanas. La información la registraron en las siguientes tablas:

Tabla 1. Nacimientos en la primera semana

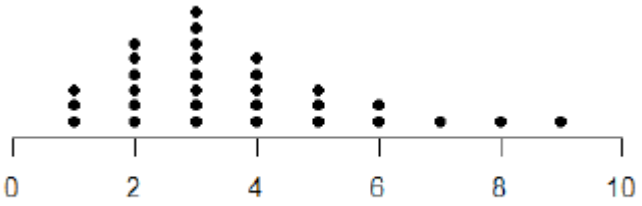
<table>
<tr><th>DÍA</th><th>HOMBRES</th><th>MUJERES</th></tr>
<tr><td>Lunes</td><td>10</td><td>8</td></tr>
<tr><td>Martes</td><td>9</td><td>13</td></tr>
<tr><td>Miércoles</td><td>7</td><td>9</td></tr>
<tr><td>Jueves</td><td>12</td><td>11</td></tr>
<tr><td>Viernes</td><td>11</td><td>8</td></tr>
<tr><td>Sábado</td><td>6</td><td>8</td></tr>
<tr><td>Domingo</td><td>9</td><td>8</td></tr>
</table>

Tabla 2. Nacimientos en la segunda semana

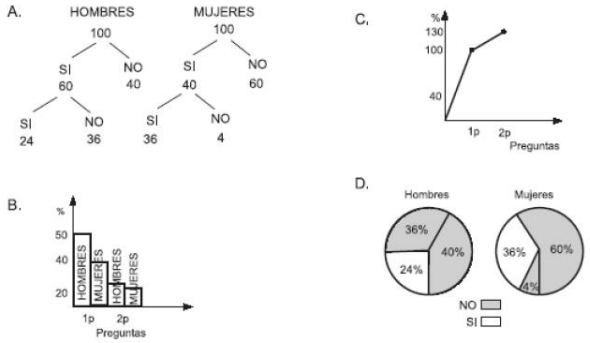
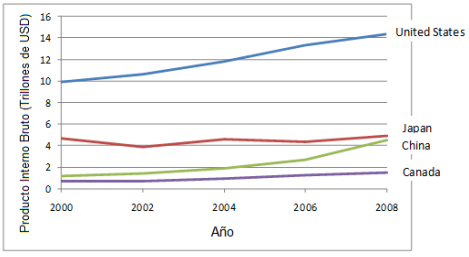
<table>
<tr><th>DÍA</th><th># TOTAL DE NACIMIENTOS</th><th>HOMBRES</th></tr>
<tr><td>Lunes</td><td>20</td><td>17</td></tr>
<tr><td>Martes</td><td>22</td><td>10</td></tr>
<tr><td>Miércoles</td><td>20</td><td>9</td></tr>
<tr><td>Jueves</td><td>18</td><td>9</td></tr>
<tr><td>Viernes</td><td>22</td><td>11</td></tr>
<tr><td>Sábado</td><td>16</td><td>4</td></tr>
<tr><td>Domingo</td><td>17</td><td>8</td></tr>
</table>

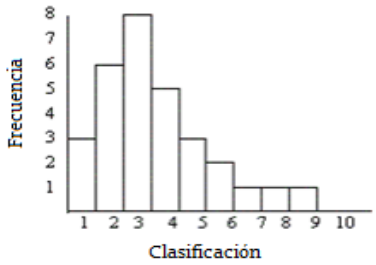
Partiendo de los datos presentados en las tablas es falso afirmar que

a. En la primera semana hubo más nacimientos que en la segunda semana
b. El nacimiento de hombres en la primera semana fue menor que el nacimiento de mujeres</td>
<td>ICFES</td>
</tr>
</table>

	c. El número de nacimientos de mujeres fue menor que el nacimiento de hombres durante las dos semanas d. El número de nacimientos de mujeres fue mayor en la segunda semana que en la primera semana	
P37	Una de las preguntas en la encuesta aplicada el primer día de clases en un curso de Estadística fue "Califique de 1 a 10 su percepción de éxito en este curso". Para responder usaron una escala donde 1 correspondía al nivel más bajo de percepción de éxito y 10 el nivel más alto. A continuación, se muestra la distribución de esta variable para los 30 estudiantes de este curso.  Calificaciones para la percepción de éxito en el curso de Estadística ¿Cómo debería el profesor interpretar las percepciones de éxito de los estudiantes en este curso? a. La mayoría de los estudiantes no sienten que ellos vayan a tener éxito en el curso, aunque unos pocos se sienten seguros de tener éxito. b. La mayoría de estudiantes en la clase califican su percepción como un 3 aunque algunas calificaciones fueron más altas y otras más bajas. c. La mayoría de estudiantes no se esforzarán por hacer bien las cosas porque ellos no sienten que puedan tener éxito en el curso. d. Un porcentaje muy bajo del grupo aprobará este curso (cerca del 17%).	ARTIST
	La siguiente gráfica muestra que la distribución de horas de sueño la noche anterior de un grupo de estudiantes de un colegio:	

P38	 Seleccione la opción que mejor describe la gráfica porque evidencia un uso adecuado de la estadística para describir gráficos y para interpretar la distribución de una variable. a. Los valores van de 3 a 10, aumentando en forma creciente hasta 7, luego decrece hasta llegar a 10. El máximo valor es 7, hay un vacío entre 3 y 5. b. La distribución es simétrica, con una media de aproximadamente 7 y una desviación estándar cercana a 1. c. Al parecer muchos estudiantes duermen 7 horas por noche, no obstante, algunos duermen más y otras menos horas. Sin embargo, un estudiante debió estar despierto gran parte de la noche por eso durmió tan pocas horas. d. La distribución de las horas de sueño es simétrica (parecida a una normal), presenta un valor extremo en 3. Un valor representativo de las horas de sueño podría ser 7 horas y la desviación estándar es cercana a 1 hora.	CAOS
P39	La empresa, Estadísticas de Colombia, realiza una encuesta a 100 hombres y 100 mujeres de Bogotá. A la 1ª pregunta responden afirmativamente el 40% de las mujeres y el 60% de los hombres. A este grupo se le hace una 2ª pregunta a la cual responden afirmativamente el 90% de las mujeres y el 40% de los hombres. Con la información suministrada por la empresa Estadística de Colombia, ¿cómo se presentarían los datos gráficamente?	ICFES

		
P40	Para determinar la clase de grafico a utilizar (histograma vs gráfico de barras) con el fin de representar una variable, el analista debe hacer cuál de las siguientes consideraciones: a. El tipo de variable b. La escala de la variable c. Si el estudio es observacional o experimental d. El rango de los datos	BLIS-1
P41	El gráfico a continuación muestra el producto interno bruto (PIB) en billones de dólares estadounidenses para cuatro países entre 2000 y 2008.  Cantidad A: El PIB combinado de Japón, China y Canadá 2008. Cantidad B: El PIB de los Estados Unidos en 2008. a. La cantidad A es mayor. b. La cantidad B es mayor. c. Las dos cantidades son iguales.	GRE

<table>
<tr><td></td><td>d. La relación no puede ser determinada a partir de la información dada.</td><td></td></tr>
<tr><td>P42</td><td>Una de las preguntas de una encuesta para un curso introductorio de estadística fue “Califique su aptitud para tener éxito en esta clase en una escala de 1 a 10” donde 1 = la aptitud más baja y 10 = la aptitud más alta. El profesor examina los datos para hombre y mujeres por separado. A continuación, se muestra la distribución de esta variable para las 30 mujeres de la clase. ¿Cómo podría el profesor interpretar la percepción femenina respecto a su éxito en esta clase?
 

a. La mayoría de mujeres en la clase no sienten que ellas tendrán éxito, aunque unas pocas se sientes confiadas sobre su desempeño.
b. Las mujeres de la clase se perciben con el más bajo nivel de aptitud para esta clase que los hombres.
c. Si usted remueve las tres mujeres con los puntajes más altos, entonces la distribución lucirá simétrica.</td><td>ARTIST</td></tr>
<tr><td>P43</td><td>Algunos estudiantes de una universidad recogieron información acerca del número de hombres y mujeres que nacieron en un hospital durante 2 semanas. La información la registraron en las siguientes tablas:

Tabla 1. Nacimientos en la primera semana

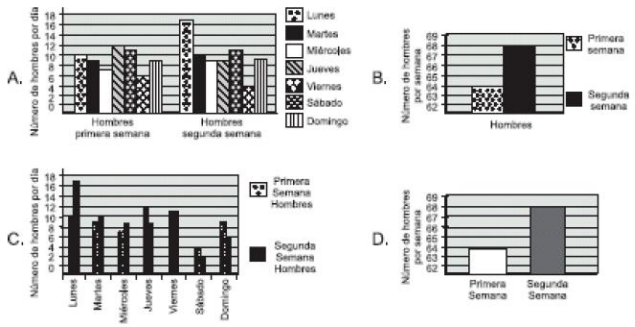
<table><tr><th>DÍA</th><th>HOMBRES</th><th>MUJERES</th></tr><tr><td>Lunes</td><td>10</td><td>8</td></tr><tr><td>Martes</td><td>9</td><td>13</td></tr><tr><td>Miércoles</td><td>7</td><td>9</td></tr></table></td><td>ICFES</td></tr>
</table>

Jueves	12	11
Viernes	11	8
Sábado	6	8
Domingo	9	8

Tabla 2. Nacimientos en la segunda semana

DÍA	# TOTAL DE NACIMIENTOS	HOMBRES
Lunes	20	17
Martes	22	10
Miércoles	20	9
Jueves	18	9
Viernes	22	11
Sábado	16	4
Domingo	17	8

Con los datos que registraron los estudiantes desean hacer una comparación entre la cantidad de hombres nacidos durante las 2 semanas. ¿Cuál de las siguientes gráficas representa mejor esta comparación?



Apéndice B. Ítems nuevos del Test Final.

7. Para dos muestras de datos sobre rendimiento académico se muestran a continuación los estadísticos descriptivos básicos y los datos, a partir de esta información seleccione la opción más adecuada para concluir sobre el rendimiento de estos dos grupos

	A	B
n	18	18
Media	58,94	59,06
Mediana	60	61,5
Moda	60	66
Desviación	9,84	10,18
Varianza	96,879	103,82
Asimetría	0,386	-0,12
Curtosis	-1,07	-1,65
Valor mínimo	46	44
Valor máximo	75	73

A	**B**
65	70
73	62
52	50
48	72
74	66
61	66
60	51
75	48
47	46
73	50
48	44
53	73
46	61
60	66
56	46
60	71
50	66
60	55

a. Hay un mejor desempeño del Grupo B (mayor promedio y mayor dispersión)
b. Rendimiento muy similar, no hay diferencias considerables
c. Mejor el grupo A, puntajes tanto mínimos como máximos son superiores
d. Mejor el grupo A, su distribución es más simétrica y baja curtosis

21. A continuación, se presentan las calificaciones del primer parcial de Estadística: 17, 13, 8, 15, 14, 11, 13, 18, 19, 20, 14, 17, 16, 18, 12, 16, 14, 15, 19, 17, 2, 4, 19, 10, 6, 18. Si un estudiante obtuvo una calificación de 14 puntos, cuál de las siguientes opciones describe mejor cómo fue su rendimiento en relación con el grupo:

a. Su rendimiento es regular pues se ubica en la media del grupo
b. No es preocupante, su nota se ubica entre el cuartil 1 y el cuartil 2
c. Rendimiento normal dado que tiene poca desviación de la nota promedio
d. Muy bueno, su nota se ubica en el centro de la distribución cerca de la mediana
e. Aceptable, su desempeño es inferior al de la mitad del grupo

Las siguientes preguntas 26-27 se responden a partir de los siguientes datos.

Los datos corresponden a una encuesta que indagó la opinión por el servicio de residencias universitarias.

	Positiva	**Negativa**	**Neutral**	Total
Hombres	424	303	131	858
Mujeres	295	675	121	1091
Total	719	978	252	1949

% Total	0,218	0,156	0,067
	0,151	0,346	0,062
% Fila	0,494	0,353	0,153
	0,27	0,619	0,111
% Columna	0,59	0,31	0,52
	0,41	0,69	0,48

26. Con relación al valor 675 de la anterior tabla, una interpretación válida es:

a. Entre las mujeres, cerca del 38% tiene una opinión negativa

b. Cerca del 33% de los encuestados son mujeres con opinión negativa

c. De la opinión negativa, el 33% corresponde a mujeres

d. El porcentaje de opinión negativa es mayor entre las mujeres debido a que hay más mujeres en la muestra

27. Cuál de las siguientes es una conclusión válida sobre la asociación entre el género y la opinión sobre el servicio de residencias (abajo aparecen algunos porcentajes calculados)

a. Hay asociación entre las variables género y opinión por ello se observa que es más probable que los hombres tengan una opinión negativa del servicio de residencias.

b. Hay asociación entre las variables género y opinión, por ello se observa que es más probable que las mujeres tengan una opinión negativa del servicio de residencias.

c. No hay asociación entre género y opinión.

d. Sólo se observa el efecto del género en la opinión negativa.

Apéndice C. Patrón de respuestas observadas.

	P2	P3	P5	P6	P7	P8	P9	P10	P12	P13	P14	P15	P16	P17	P18	P19	P20	P21	P22	P23	P24	P25	P26
1	0	0	0	0	0	0	0	1	1	0	1	1	1	0	0	1	0	0	0	0	0	1	0

2	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	1	1	0	1	1	0	1	0
3	0	0	0	0	0	1	0	1	0	0	1	1	1	1	0	1	0	0	0	0	1	0	1
4	0	0	0	0	0	1	1	0	1	1	1	1	0	0	0	0	0	0	1	1	1	0	0
5	0	0	1	0	0	0	0	0	0	0	0	1	0	0	1	0	1	0	0	0	0	1	1
6	0	0	1	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	1	0	0	1	1
7	0	0	1	0	0	0	0	1	1	1	1	0	1	0	0	0	0	0	0	0	0	1	1
8	0	0	1	0	0	1	0	1	1	1	1	0	0	0	0	0	0	0	1	0	0	0	1
9	0	0	1	1	0	1	1	1	1	0	1	1	1	0	1	1	0	0	0	1	0	0	1
10	0	0	1	1	0	1	1	1	1	0	1	1	1	1	0	1	1	0	1	1	1	0	1
11	0	0	1	1	1	0	1	1	1	1	1	1	1	0	0	1	1	0	0	0	1	0	1
12	0	1	0	0	0	1	0	1	1	0	0	1	1	1	1	0	1	0	0	0	1	0	0
13	0	1	0	0	0	1	1	0	0	0	0	0	0	1	0	1	1	0	0	1	0	1	0
14	0	1	0	0	1	0	0	0	0	0	1	1	1	0	0	0	0	0	1	1	1	1	1
15	0	1	0	0	1	1	0	0	0	0	0	1	1	0	0	0	0	0	1	1	0	0	1
16	0	1	0	0	1	1	0	1	0	0	1	0	1	1	1	0	0	1	0	0	0	1	1
17	0	1	0	0	1	1	0	1	0	1	0	1	1	1	0	0	1	0	1	0	1	0	1
18	0	1	0	0	1	1	0	1	1	1	0	1	1	0	0	0	1	0	0	1	1	0	1
19	0	1	0	1	0	1	0	0	0	0	0	0	0	1	0	1	1	0	0	0	1	0	1
20	0	1	0	1	0	1	0	0	0	0	0	1	1	1	1	1	0	1	0	1	0	0	1
21	0	1	0	1	0	1	0	1	1	0	1	1	1	0	1	1	1	0	1	1	1	0	1
22	0	1	0	1	0	1	1	1	0	0	1	1	1	0	1	1	1	1	0	1	0	0	1
23	0	1	0	1	1	1	1	1	0	0	0	1	1	1	0	0	0	0	1	1	0	0	0
24	0	1	1	0	0	0	0	1	0	0	0	1	1	0	1	1	1	0	0	1	0	1	1
25	0	1	1	0	0	1	0	1	0	0	0	1	1	0	0	0	0	0	1	1	1	0	1
26	0	1	1	0	0	1	0	1	0	0	0	1	1	0	0	1	0	0	1	1	0	0	1
27	0	1	1	0	0	1	1	1	1	1	1	1	1	0	1	0	0	0	0	1	0	1	1
28	0	1	1	0	0	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1	0	0	0
29	0	1	1	0	1	1	0	0	1	1	1	1	1	0	1	1	1	1	0	1	0	0	1
30	0	1	1	0	1	1	0	1	0	0	1	0	1	1	1	1	0	0	0	1	0	0	1
31	0	1	1	0	1	1	0	1	0	0	1	1	1	1	1	1	0	0	1	1	1	0	1
32	0	1	1	0	1	1	1	0	1	0	1	1	1	1	0	1	1	1	1	1	1	1	0
33	0	1	1	0	1	1	1	1	1	0	1	1	1	0	1	1	1	0	1	1	0	1	1
34	0	1	1	0	1	1	1	1	1	1	1	0	0	0	0	1	1	0	1	1	1	1	1
35	0	1	1	0	1	1	1	1	1	1	1	1	0	0	1	0	0	0	1	1	1	1	1
36	0	1	1	1	0	0	0	1	1	1	1	0	0	0	0	0	0	0	1	0	1	0	0
37	0	1	1	1	0	1	0	1	0	0	1	1	0	1	1	0	0	0	1	1	0	1	1
38	0	1	1	1	0	1	0	1	1	0	0	1	0	0	1	0	0	0	1	1	1	0	0
39	0	1	1	1	0	1	0	1	1	0	0	1	1	0	1	1	0	1	1	1	1	0	1
40	0	1	1	1	0	1	0	1	1	0	0	1	1	1	1	0	0	0	1	1	1	0	1
41	0	1	1	1	0	1	0	1	1	0	1	1	1	0	1	0	0	0	0	1	0	1	1
42	0	1	1	1	0	1	1	1	0	0	1	1	0	1	0	1	1	0	1	1	1	1	0
43	0	1	1	1	0	1	1	1	0	0	1	1	1	0	1	0	0	0	0	1	1	0	1

44	0	1	1	1	1	1	0	0	0	1	0	0	1	1	0	1	0	1	1	1	1	0	0
45	0	1	1	1	1	1	0	1	0	0	0	1	0	0	1	1	1	0	1	1	1	0	1
46	0	1	1	1	1	1	0	1	0	0	1	1	0	1	1	1	0	0	1	1	0	0	1
47	0	1	1	1	1	1	0	1	1	1	1	1	0	1	0	1	0	0	1	1	0	1	1
48	0	1	1	1	1	1	1	0	1	1	1	1	0	0	0	1	0	1	0	0	0	1	1
49	0	1	1	1	1	1	1	1	1	0	1	1	1	1	0	1	0	0	1	1	0	1	1
50	0	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	0	0	1	1	1	1	1
51	1	0	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	0	1	1	0	0	1
52	1	1	0	1	1	1	0	1	1	0	1	1	1	0	0	0	1	0	1	1	1	0	1
53	1	1	0	1	1	1	1	0	1	0	1	1	1	0	1	1	1	0	1	1	1	0	1
54	1	1	1	0	0	1	1	1	1	0	1	1	0	0	1	1	0	0	0	1	0	1	1
55	1	1	1	0	1	1	0	1	0	0	0	1	1	0	1	1	1	0	1	0	0	0	1
56	1	1	1	0	1	1	1	0	1	1	1	1	1	1	1	1	0	0	0	1	1	1	1
57	1	1	1	1	0	1	0	1	1	0	1	1	0	0	0	0	0	0	1	1	0	0	1
58	1	1	1	1	0	1	0	1	1	1	1	1	1	0	0	1	1	0	1	1	0	0	1
59	1	1	1	1	0	1	1	1	0	0	0	1	1	0	0	0	0	0	1	1	0	0	1
60	1	1	1	1	1	1	0	1	1	0	1	1	0	1	1	1	0	1	0	1	1	1	1
61	1	1	1	1	1	1	0	1	1	1	1	1	1	0	0	0	0	0	1	1	1	1	1
62	1	1	1	1	1	1	0	1	1	1	1	1	1	0	1	1	1	0	0	1	0	0	1
63	1	1	1	1	1	1	1	1	0	0	1	1	1	1	0	0	0	0	1	1	1	1	1
64	1	1	1	1	1	1	1	1	1	0	1	1	1	0	0	1	1	1	1	1	1	1	1
65	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1
66	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1

Apéndice D. Rutinas en R.

D.1 Rutina en R de la Unidimensionalidad por medio del Análisis Factorial.

AF1

Call:

factanal(x = datosirtmodel, factors = 1)

Uniquenesses:

P1 P2 P3 P4 P5 P6 P7 P8 P9 P10 P11 P12 P13

0.992 0.751 0.754 0.980 0.885 0.805 0.845 0.651 0.920 0.932 0.997 0.891 0.993

P14 P15 P16 P17 P18 P19 P20 P21 P22 P23 P24 P25 P26

0.873 0.829 0.973 0.999 0.899 0.956 0.996 0.972 0.948 0.634 0.991 1.000 0.902

P27

0.991

Loadings:

Factor1

P1

P2 0.499

P3 0.496

P4 0.140

P5 0.339

P6 0.441

P7 0.394

P8 0.591

P9 0.282

P10 0.262

P11

P12 0.330

P13

P14 0.356

P15 0.414

P16 0.166

P17

P18 0.317

P19 0.210

P20

P21 0.167

P22 0.229

P23 0.605

P24

P25

P26 0.313

P27

Factor1

SS loadings 2.642

Proportion Var 0.098

Test of the hypothesis that 1 factor is sufficient.

The chi square statistic is 371.54 on 324 degrees of freedom.

The p-value is 0.0352

D.2 Rutinas en R del modelo de un parámetro (1P)

```
install.packages("ltm")
library(ltm)
datosirtmodel<-read.table("(dirección del archivo)",header=TRUE)
mod.1pl <-rasch(datosirtmodel)
mod.1pl$conv
coef(mod.1pl)
factor.scores(mod.1pl)
```

```
person.fit(mod.1pl)
item.fit(mod.1pl)
plot(mod.1pl,type="ICC",items=c(1,2,3,4))
plot(mod.1pl,type="IIC",items=0)
```

D.3 Rutinas en R del modelo de dos parámetros (2P).

```
install.packages("ltm")
library(ltm)
datosirtmodel<-read.table("(dirección del archivo)",header=TRUE)
modelo2=ltm(datosirtmodel~z1,IRT.param=TRUE)
modelo2
factor.scores(modelo2)
person.fit(modelo2)
item.fit(modelo2)
plot(modelo2,type="ICC",items=c(1,2,3,4))
plot(modelo2,type="IIC",items=0)
```

D.4 Rutinas en R del modelo de tres parámetros (3P)

```
install.packages("ltm")
library(ltm)
datosirtmodel<-read.table("(dirección del archivo)",header=TRUE)
modelo3=tpm(datosirtmodel,type="latent.trait",IRT.param=TRUE)
modelo3
factor.scores(modelo3)
```

```
person.fit(modelo3)

item.fit(modelo3)

plot(modelo3,type="ICC",items=c(1,2,3,4))

plot(modelo3,type="IIC",items=0)
```

D.5 Rutinas en R de comparación de los modelos

```
anova(mod.1pl,modelo2)

anova(mod.1pl,modelo3)

anova(modelo2,modelo3)
```